



Dr. Vishwanath Karad

**MIT WORLD PEACE
UNIVERSITY** | PUNE

TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS

MACHINE LEARNING

Professional Core(CET3006B)

T. Y. B.Tech CSE, Sem-VI
2023-2024

SoCSE – Dept. of Computer Engineering & Technology



MACHINE LEARNING

- **Credits :** 3+1 (Four)
- **Examination scheme:** Total Marks-100
 - 30 Marks CCA
 - 30 Marks LCA
 - 40 Marks End Term Examination

MACHINE LEARNING



Course Objectives:

1.Knowledge:

- i. To learn data preparation techniques for Machine Learning methods
- ii. To understand advance supervised and unsupervised Learning methods

2.Skills:

- i. To apply suitable pre-processing techniques on various datasets for Machine Learning applications.
- ii. To design and implement various advanced supervised and unsupervised learning methods

3.Attitude:

- i. To be able to choose and apply suitable ML techniques to solve the problem
- ii. To compare & analyze various advanced supervised and unsupervised learning methods.



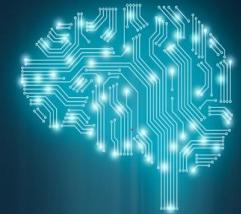
ADVANCES IN MACHINE LEARNING

Course Outcomes:

After completion of the course the students will be able to:

1. Analyze and apply different data preparation techniques for Machine Learning applications
2. Identify, Analyze and compare appropriate supervised learning algorithm for given problem
3. Identify, Analyze and Compare Unsupervised and semi supervised algorithms
4. Design and implement Machine Learning techniques for real-time applications

Course Contents:



Unit 1. Introduction to ML

Unit 2. Supervised Learning: Classification

Unit 3. Unsupervised Learning: Clustering

Unit 4. Performance Analysis and Model Evaluation

Unit 5. Trends in ML

Course Contents:



Laboratory Exercises:

1. Implement various Pre-processing techniques on given dataset.
2. Implement KNN classifier for given dataset.
3. Implementation of Tree based Classifiers.
4. Implementation of SVM, Comparison with Tree Based Classifier
5. Implementation of Ensemble, Random Forests. Analyze the Performance.
6. Implementation and Comparison of various clustering techniques such as Spectral & DBSCAN.
7. Implement Regression Technique & evaluate its performance.
8. Mini-Project based on suitable Machine Learning dataset



Course Contents:

Text Books:

1. E. Alpaydin, Introduction to Machine Learning, PHI, 2004.
2. Peter Flach: Machine Learning: The Art and Science of Algorithms that Make Sense of Data, Cambridge University Press, Edition 2012.
3. T. Mitchell, Machine Learning, McGraw-Hill, 1997.
4. Josh Patterson, Adam Gibson, “Deep Learning: A Practitioners Approach”, O“REILLY, SPD, ISBN: 978-93-5213-604-9, 2017 Edition 1st

Reference Books:

1. C. M. Bishop: Pattern Recognition and Machine Learning, Springer 1st Edition-2013.
2. Ian H Witten, Eibe Frank, Mark A Hall: Data Mining, Practical Machine Learning Tools and Techniques, Elsevier, 3rd Edition.
3. Shaishalev-shwartz, Shai Ben-David: Understanding Machine Learning from Theory to algorithms, Cambridge University Press, ISBN-978-1-107-51282-5, 2014.

Course Contents:



Supplementary Reading:

1. AurelienGeron, “Hands-on Machine Learning with Scikit-learn and Tensor flow, O’Reilly Media

Web Resources:

1. Popular dataset resource for ML beginners: <http://archive.ics.uci.edu/ml/index.php>

Web links:

1. <https://www.kaggle.com/datasets>
2. <http://deeplearning.net/datasets/>

MOOCs:

1. https://swayam.gov.in/nd1_noc20_cs29/preview
2. https://swayam.gov.in/nd1_noc20_cs44/preview



Course Contents:

Assessment Scheme:

Class Continuous Assessment (CCA): 30 Marks

Mid Term	Component 1 (Active Learning)	Component 2
15 Marks	10 Marks	5 Marks

Laboratory Continuous Assessment (LCA): 30 Marks

Practical Performance	Active learning / Mini Project/Additional implementation/ On paper design	End term practical /oral examination
10 Marks	10 Marks	10 Marks

Term End Examination: 40 Marks



Syllabus-Unit 1

Introduction to ML:

Introduction, Data Preparation

Data Encoding Techniques

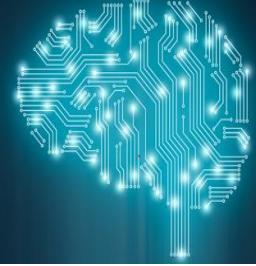
Data Pre-processing techniques for ML applications.

Feature Engineering:

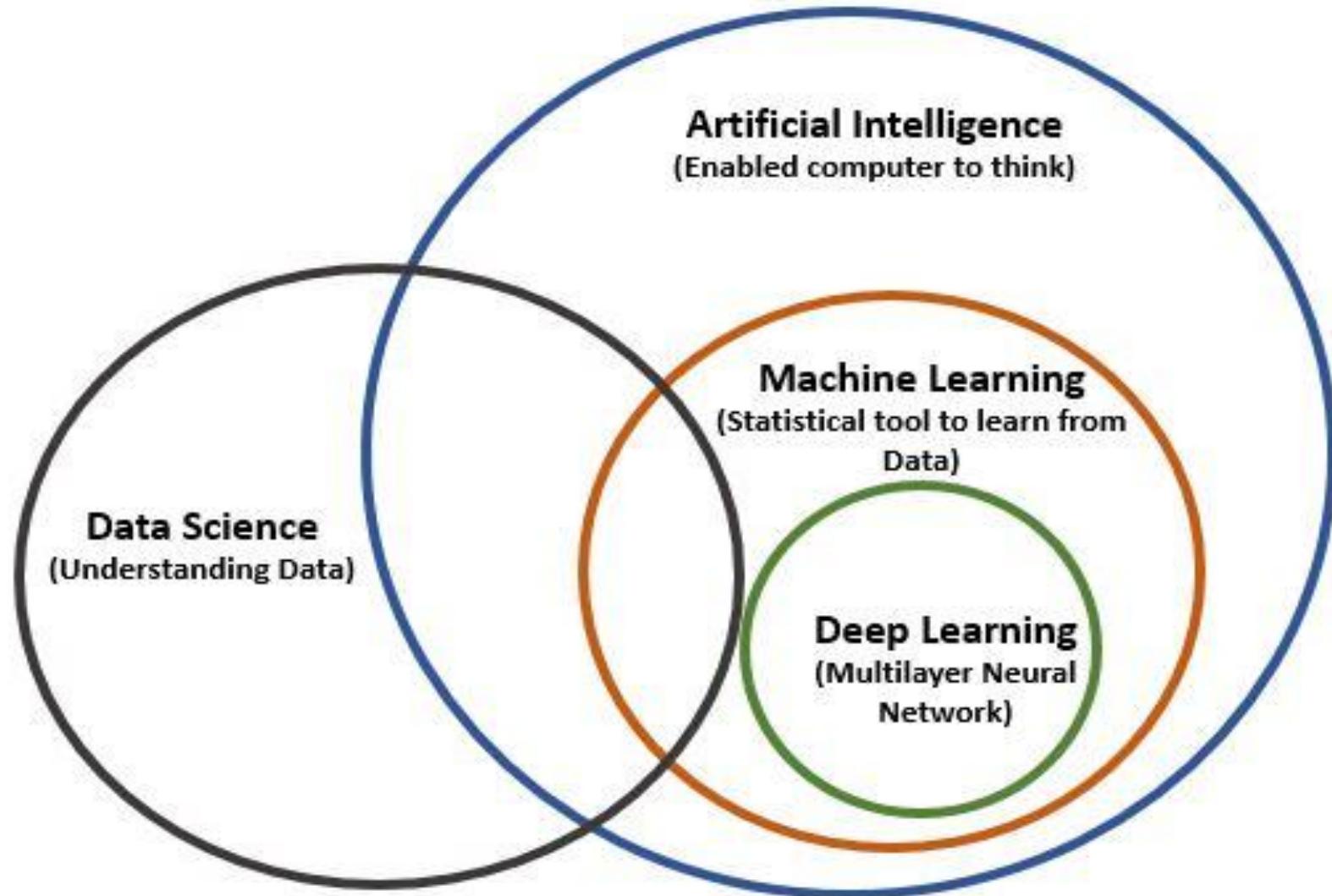
Dimensionality Reduction using PCA

Exploratory Data Analysis

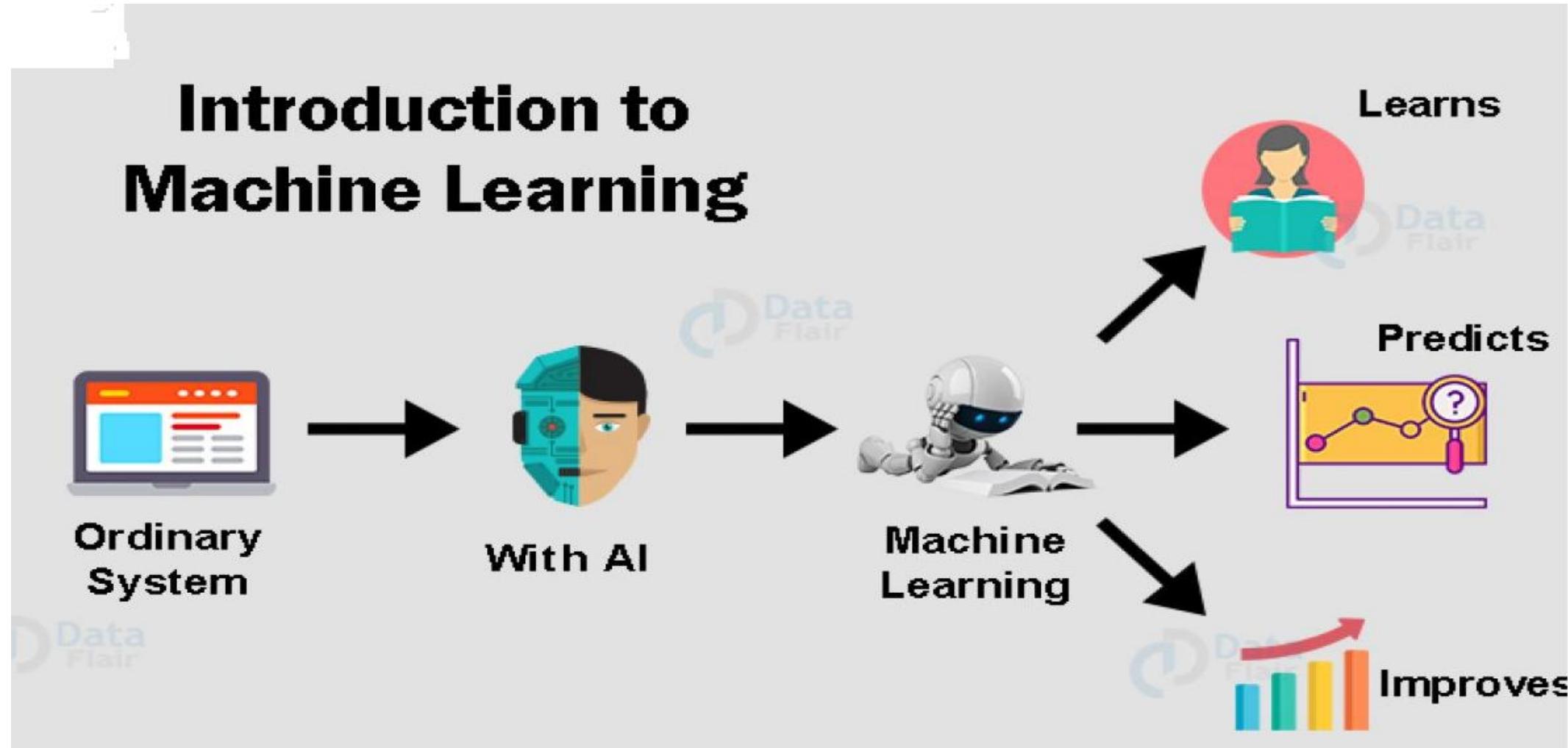
Feature Selection



AI Vs. ML



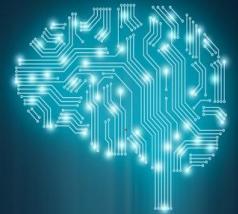
INTRODUCTION





MIT-WPU

॥ विद्यशान्तिर्धूमं ध्रुवा ॥



Cntd..

To solve a problem on a computer, we need an algorithm.

An algorithm is a sequence of instructions that should be carried out to transform the input to output

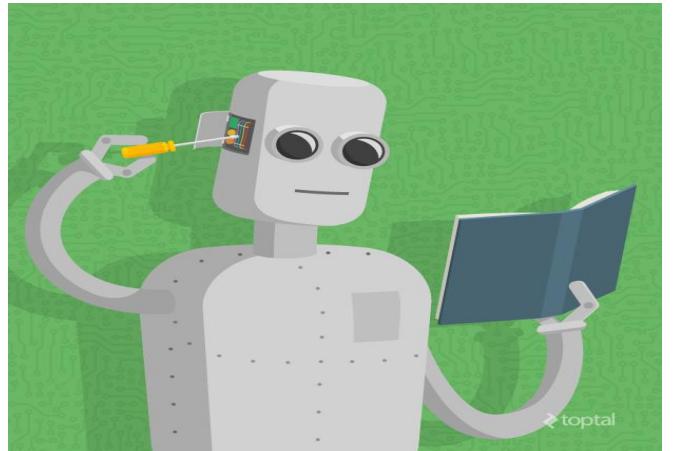
Ex. Sorting Input :set of numbers , output : ordered list

For some tasks ,however, we do not have an algorithm We are machine learning intelligence

Ex. tell spam emails from legitimate emails

Input :email document (file of characters), output : yes/no output indicating whether the message is spam or not

computer (machine) to extract automatically the algorithm for this task.

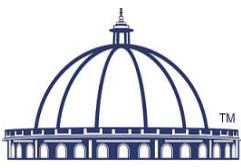


Traditional Programming



Machine Learning

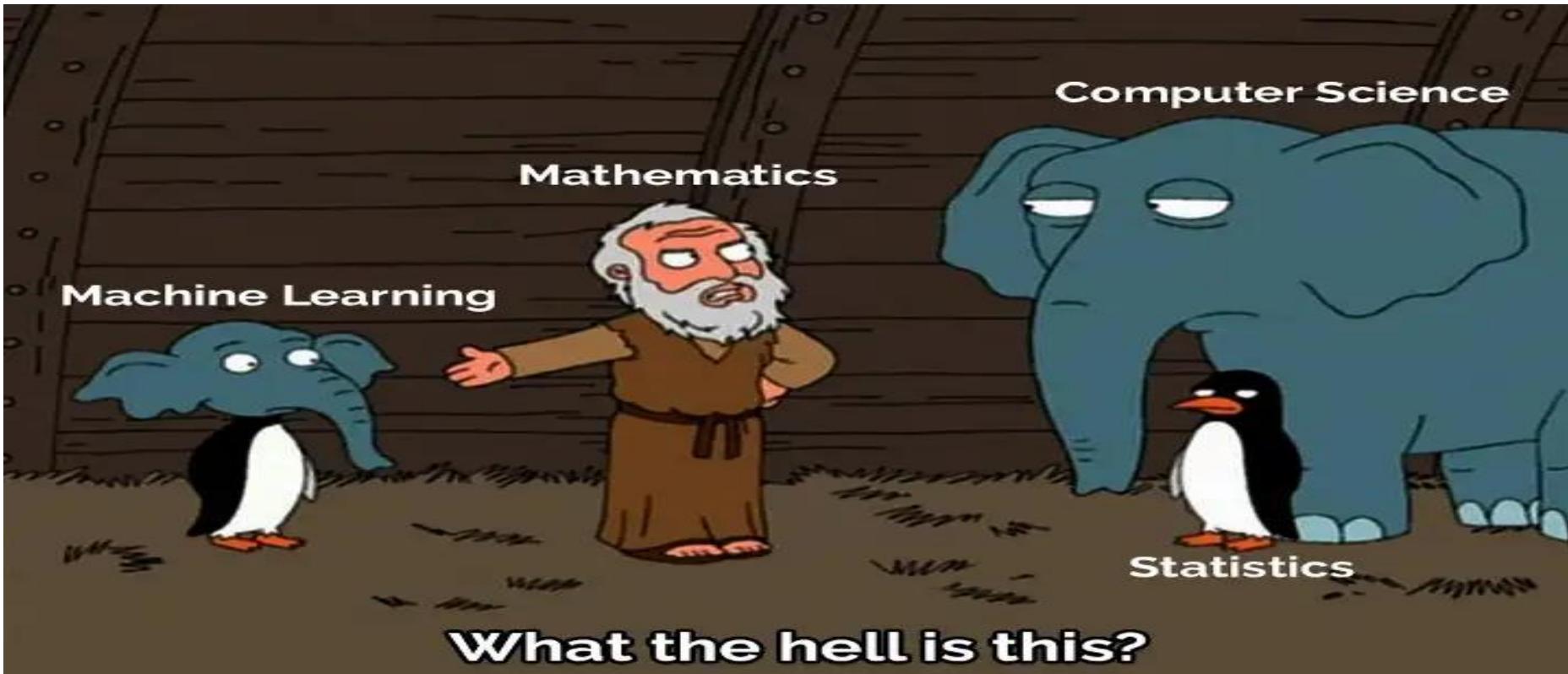




MIT-WPU

॥ विद्यानन्तर्धारा ॥

Cntd..



(source: <https://medium.com/analytics-vidhya/introduction-to-machine-learning-e1b9c055039c>)

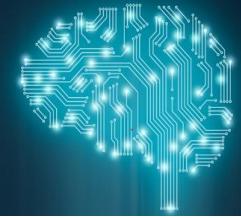
- Machine learning is a “***Field of study that gives computers the ability to learn without being explicitly programmed.***”
- In other words it is concerned with the question of ***how to construct computer programs that automatically improve with the experience.*** - According to Arthur Samuel(1959)



Cntd..

- A **computer program** is said to **learn from experience ‘E’** with respect to some class of **task ‘T’** and **performance measure ‘P’** if its performance at task in ‘T’ as measured by ‘P’ **improves with experience ‘E’** – Tom M Mitchell
- Machine learning is an application of artificial intelligence (AI) that provides systems the **ability to automatically learn and improve from experience** without being explicitly programmed.
- Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

Cntd..



Example 1

Classify Email as spam or not spam

- Task (T): Classify email as spam or not spam
- Experience(E): watching the user to mark/label the email as spam or not spam
- Performance (P): The number or fraction of email to be correctly classified as spam or not spam



Cntd..

Example 2

Recognizing hand written digits/ characters

- Task(T): Recognizing hand written digit
- Experience (E): watching the user to mark/ label the hand written digit to 10 classes(0-9) & identify underling pattern
- Performance(P):The number of fractions of hand-written digits correctly classified



Why Machine Learning Important?.

- **Human expertise does not exist**

Navigating on Mars

industrial/manufacturing control

mass spectrometer analysis, drug design, astronomic discovery

- **Black-box human expertise OR Some tasks cannot be defined well, except by examples**

face/handwriting/speech recognition/ recognizing people

driving a car, flying a plane

- **Relationships and correlations can be hidden within large amounts of data**

(e.g., stock market analysis)

- **Environments change over time.**

(e.g., routing on a computer network)



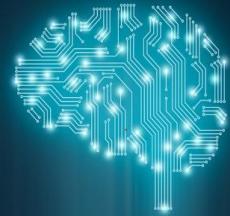
MIT-WPU

॥ विद्यानन्तर्धावं ध्रुवा ॥



Cntd..

- **The amount of knowledge available about certain tasks might be too large for explicit encoding by humans**
(e.g., medical diagnostic).
- **New knowledge about tasks is constantly being discovered by humans. It may be difficult to continuously re-design systems “by hand”.**
- **Rapidly changing phenomena**
credit scoring, financial modeling
diagnosis, fraud detection
- **Need for customization/personalization**
personalized news reader
movie/book recommendation



How does Machine Learning help us in daily life?

Social networking :

- Use of the appropriate emotions, suggestions about friend tags on Facebook, filtered on Instagram, content recommendations and suggested followers on social media platforms, etc., are examples of how machine learning helps us in social networking.

Personal finance and banking solutions

- Whether it's fraud prevention, credit decisions, or checking deposits on our smartphones machine learning does it all.

Commute estimation

- Identification of the route to our selected destination, estimation of the time required to reach that destination using different transportation modes, calculating traffic time, and so on are all made by machine learning.



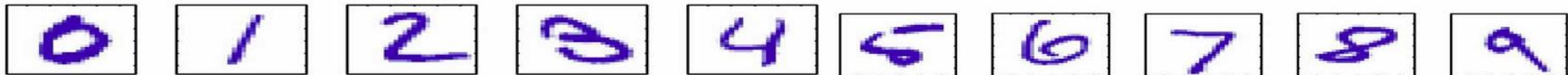
Applications of Machine Learning

- Face detection
- Stock prediction
- Spam Email Detection
- Machine Translation
- Self-parking Cars
- Airplane Navigation Systems
- Medicine
- Data Mining
- Speech recognition
- Hand-written digit recognition
- Computational Biology
- Recommender Systems
- Guiding robots
- Space Exploration
- Supermarket Chain

Examples...



Example 1: hand-written digit recognition: **Output**



Learn a classifier $f(x)$ such that, $f : x \rightarrow \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$

Input training data : e.g. 500 samples

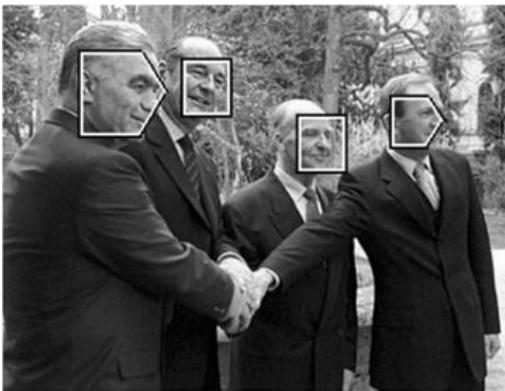
0 0 0 1 1 1 1 1 1 2
2 2 2 2 2 2 2 3 3 3
3 4 9 9 9 4 5 5 5 5
6 6 7 7 7 7 7 8 8 8
8 8 9 9 9 4 9 9 9

Example 2: Face detection

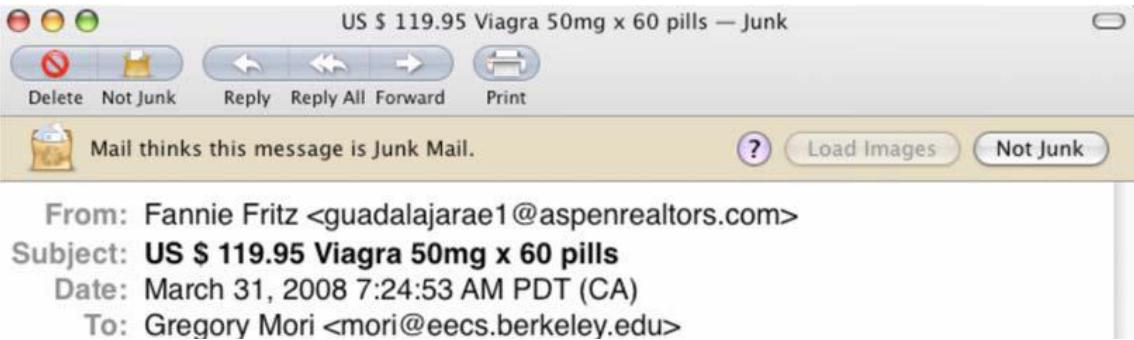
Input : an **image**, the classes are people to be recognized ...

[**non-face**, **frontal-face**, **profile-face**] and the learning program should learn to associate the face images to identities.

This problem is more difficult because there are **more classes**, **input image is larger**, and a face is **3-dimensional** and differences in pose and lighting cause significant changes in the image. There may also be **occlusion** (blockage) of certain inputs; e.g. glasses may hide the eyes and eyebrows, and a beard may hide the chin.



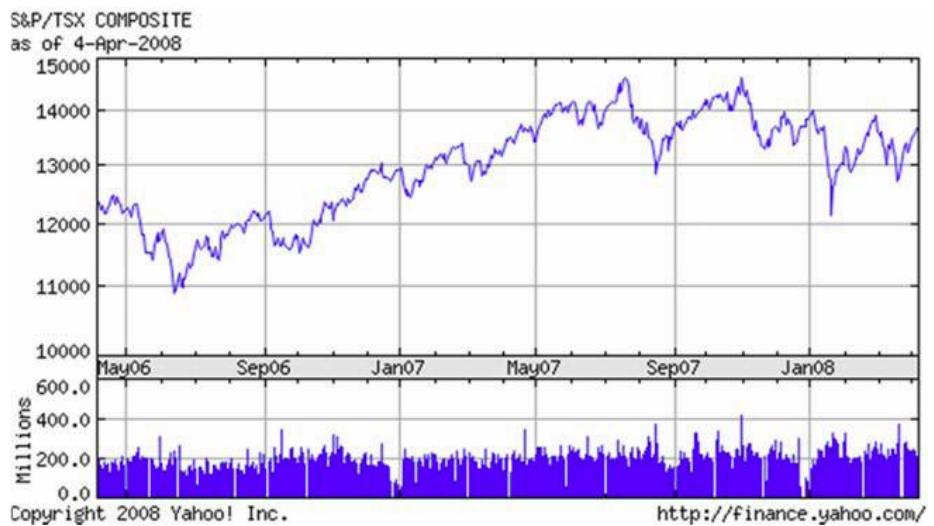
Example 3: Spam detection



- This is a classification problem
- Task is to classify email into spam/non-spam
- Requires a learning system as “enemy” keeps innovating



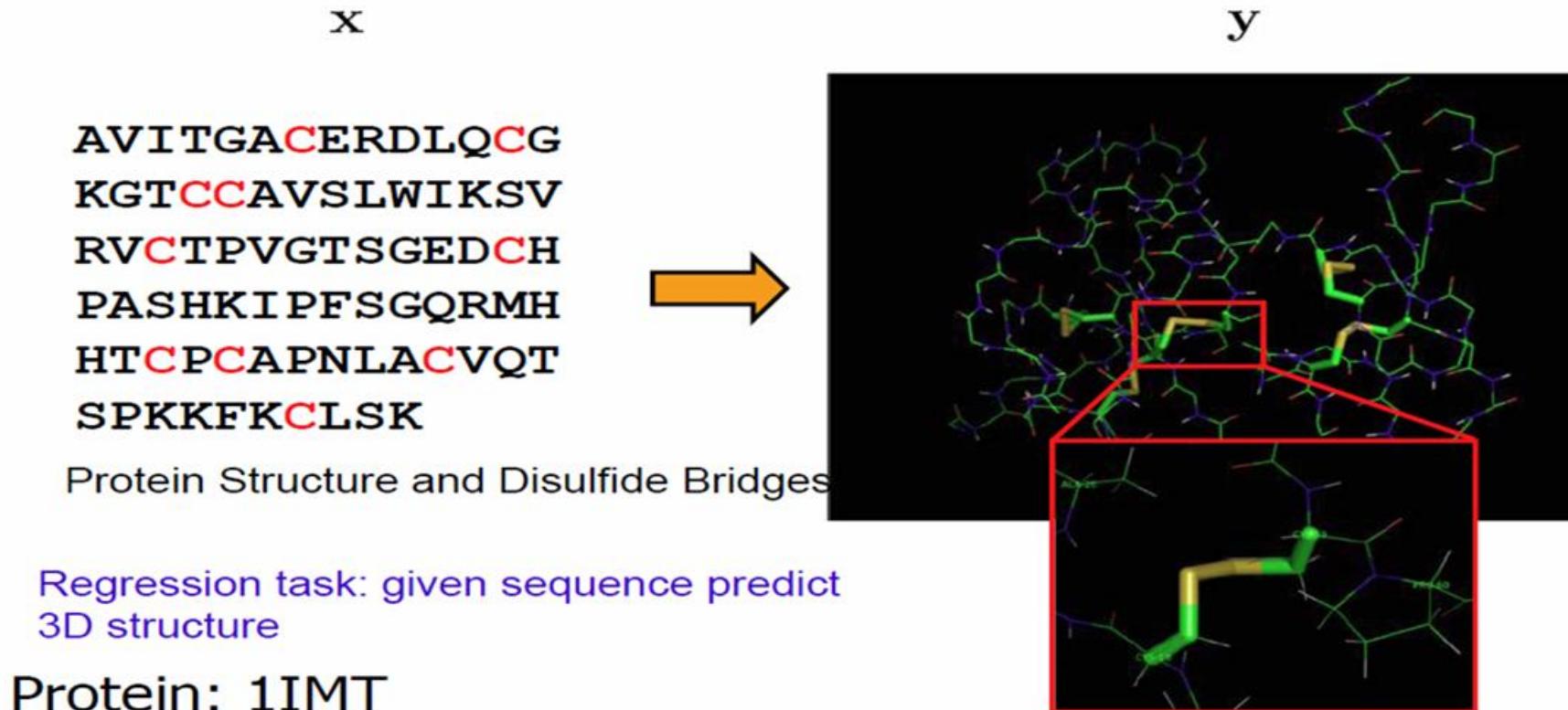
Example 4: Stock price prediction

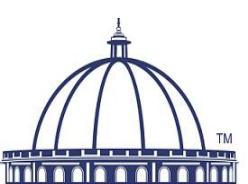


- Task is to predict stock price at future date
- This is a regression task, as the output is continuous



Example 5: Computational Biology





MIT-WPU

॥ विद्यशान्तिर्धूमं ध्रुवा ॥



lect1.pdf - Adobe Reader

File Edit View Document Tools Window Help

15 (7 of 24) 132% Find

Web examples: Machine translation

Use of aligned text

X

What is the anticipated cost of collecting fees under the new proposal?

En vertu des nouvelles propositions, quel est le coût prévu de perception des droits?

Y

En vertu de les nouvelles propositions, quel est le coût prévu de perception des droits?

What is the anticipated cost of collecting fees under the new proposal?

e.g. Google translate

2:09 PM
6/19/2017

Web Images Maps News Shopping Mail more ▾

Help

Google Translate BETA

Home Text and Web Translated Search Dictionary Tools

Translate text or webpage

Enter text or a webpage URL.

Translation: French » English

En vertu des nouvelles propositions, quel est le coût prévu de perception des droits?

Under the new proposals, what is the cost of collection of fees?

French ▾ > English ▾ swap Translate

Suggest a better translation



MIT-WPU

॥ विद्यानन्तर्धारा ॥



lect1.pdf - Adobe Reader

File Edit View Document Tools Window Help

17 (8 of 24) 132% Find

What is the anticipated cost of collecting fees under the new proposal?

Web examples: Recommender systems

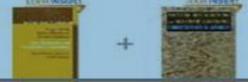
People who bought Hastie ...

Frequently Bought Together

Customers buy this book with [Pattern Recognition and Machine Learning \(Information Science and Statistics\)](#) (Information Science and Statistics) by Christopher M. Bishop

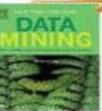
Price For Both: £104.95

Add both to Basket



Customers Who Bought This Item Also Bought

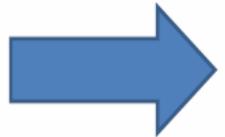
Page 1

 Pattern Recognition and Machine Learning (Infor... View Details	 MACHINE LEARNING (Mcgraw-Hill International) View Details	 Pattern Classification, Second Edition: 1 (A Wi... View Details	 Data Mining: Practical Machine Learning Tools a... View Details
--	--	--	--

2:12 PM 6/19/2017



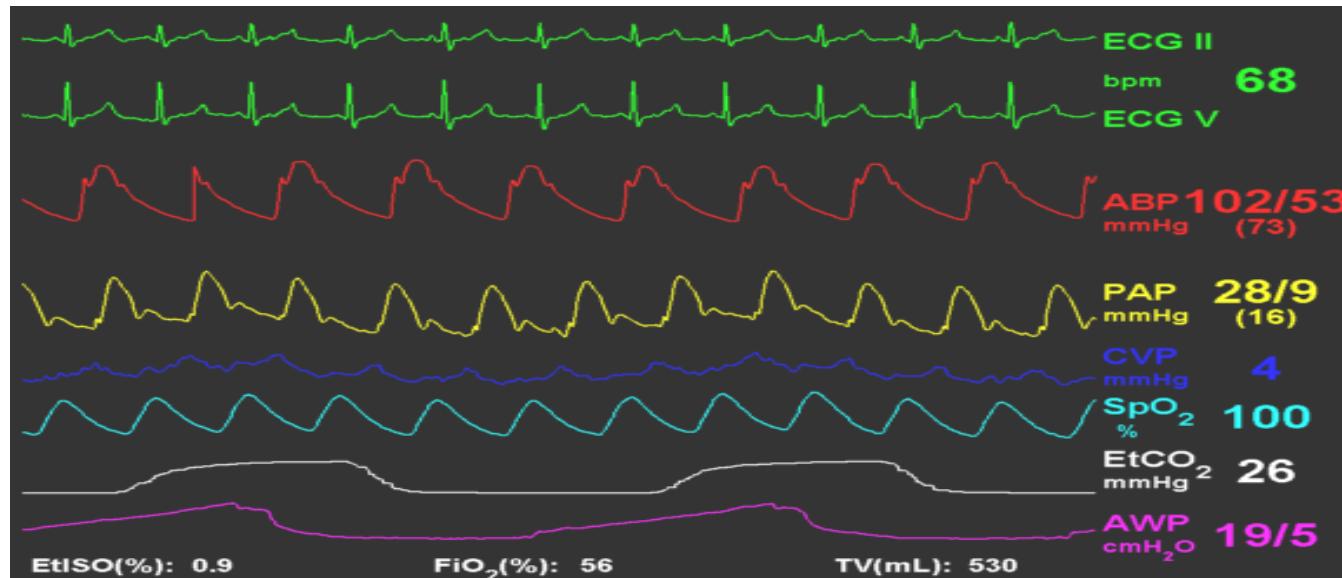
Example : Weather prediction





Example : Medical Diagnosis

- ❖ Inputs are the relevant **information** about the patient and the classes are the illnesses.
- ❖ The inputs contain the **patient's age, gender, past medical history, and current symptoms**.
- ❖ Some tests may not have been applied to the patient, and thus these inputs would be missing.
- ❖ **Tests** take time, may be costly, and may inconvenience the patient so we **do not want to apply** them unless we **believe that they will give us valuable information**.
- ❖ In the case of a medical diagnosis, a wrong decision may lead to a wrong or no treatment, and in cases of doubt it is preferable that the classifier reject and defer decision to a human expert.





Example : Agriculture



A Crop Yield Prediction App in Senegal Using Satellite Imagery (Video Link)

<https://www.youtube.com/watch?v=4OnBGkhA4jc&t=160s>

.



Data Preparation

Data Preparation Pipeline

Data Preparation

1. Clean Data

- a. do a preliminary exploration of data
- b. deal with spurious lines
- c. find and deal with missing data
- d. remove unwanted columns
- e. find and deal with bad numeric data
- f. find and deal with bad categorical data

2. Normalize and Encode Data

- a. normalize numeric predictors
- b. encode categorical predictors
- c. encode categorical dependent variable

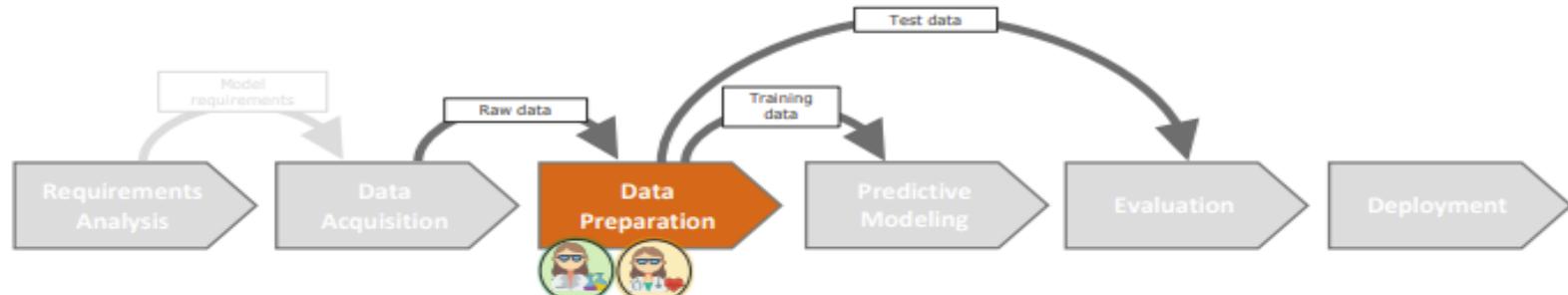
3. Split Data

- a. split data into train and test sets
- b. split train and test into train_x, train_y, test_x, test_y



Data Preparation

- Data preparation is defined as a **gathering, combining, cleaning, and transforming raw data** to make accurate predictions or uncover insights in Machine learning projects.



Roles



Data Scientist



Domain Expert

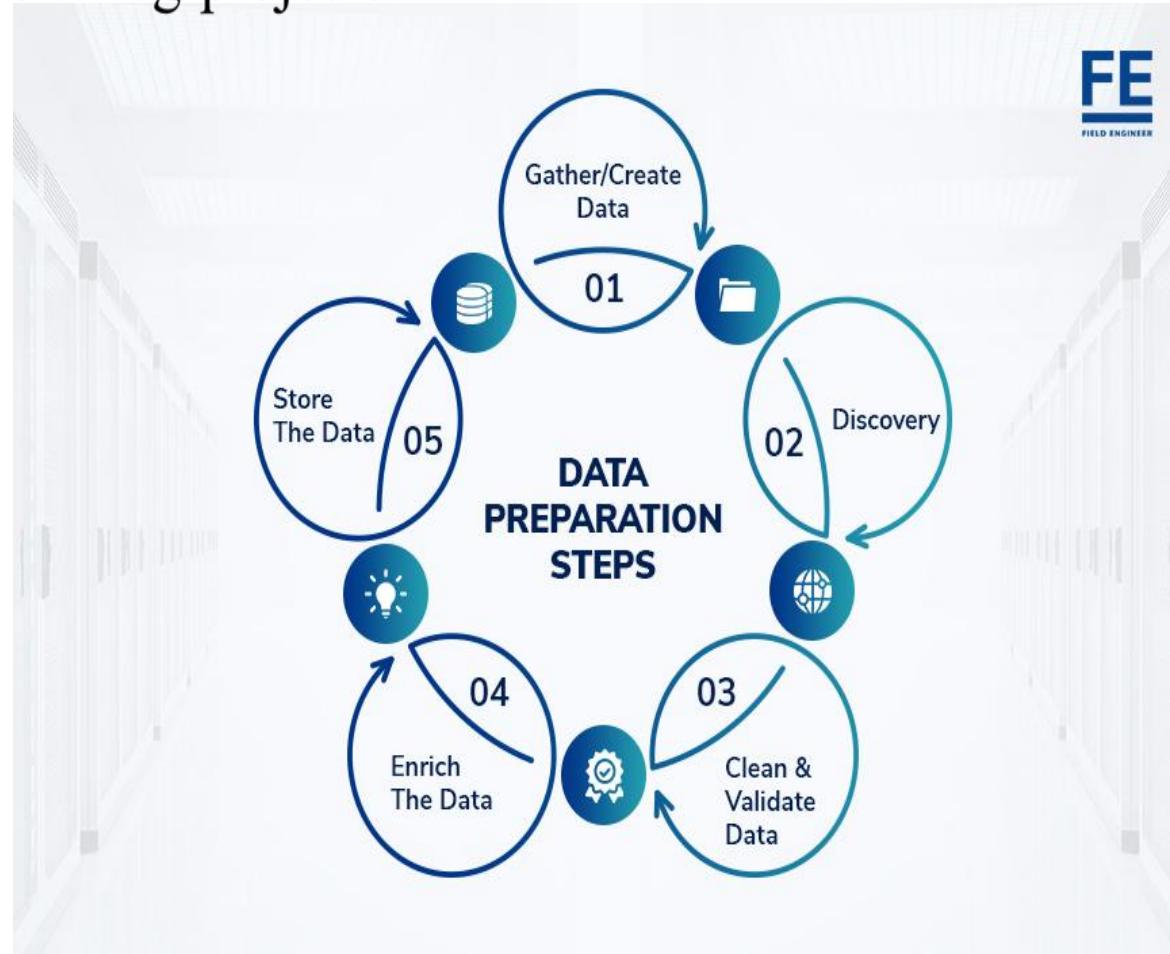


(Data) Engineer

Data Preparation



- Data preparation is defined as a **gathering, combining, cleaning, and transforming raw data** to make accurate predictions or uncover insights in Machine learning projects.





Why is Data Preparation important?

॥ विद्यशान्तिर्धूवं ध्रुवा ॥

sometimes, data in data sets **have missing or incomplete information**, which **leads to less accurate or incorrect predictions**.

Further, sometimes data sets are clean **but not adequately shaped**, such as aggregated or pivoted, and some have less business context.

Hence, after collecting data from various data sources, **data preparation needs to transform raw data**.

Significant advantages of data preparation in machine learning as follows:

- It helps to provide **reliable prediction outcomes** in various analytics operations.
- It helps **identify data issues or errors** and significantly reduces the chances of errors.
- It increases **decision-making capability**.
- It **reduces overall project cost** (data management and analytic cost).
- It helps to **remove duplicate content** to make it worthwhile for different applications.
- It **increases model performance**.



Steps in Data Preparation Process

1. Understand the problem:

Understand the actual problem and try to solve it.

2. Data collection:

collect data from various potential sources. These data sources may be either **within enterprise or third parties vendors**.

Data collection is beneficial to **reduce and mitigate biasing** in the ML model.

So, before collecting data, always analyze it and also ensure that the **data set was collected from diverse people, geographical areas, and perspectives**.

3. Profiling and Data Exploration:

explore data such as trends, outliers, exceptions, incorrect, inconsistent, missing, or skewed information, etc.

Data exploration helps to determine problems such as **collinearity**, which means a situation when the Standardization of data sets and other data transformations are necessary.



Steps in Data Preparation Process

4. Data Cleaning and Validation:

Data cleaning and validation techniques help determine and solve inconsistencies, outliers, anomalies, incomplete data, etc.

Clean data helps to find valuable patterns and information in data and ignores irrelevant data in the datasets.

5. Data Formatting:

After cleaning and validating data, the following approach is to ensure that the data is correctly formatted or not.



Steps in Data Preparation Process

6. Feature engineering and selection:

- Feature engineering is defined as the study of selecting, manipulating, and transforming raw data into valuable features

There are various feature engineering techniques used in Machine Learning as follows:

Imputation:

- Feature imputation is the technique to fill incomplete fields in the datasets.
- It is essential because most machine learning models don't work when there are missing data in the dataset.
- Although, the missing values problem can be reduced by using techniques such as single value imputation, multiple value imputation, K-Nearest neighbor, deleting the row, etc.

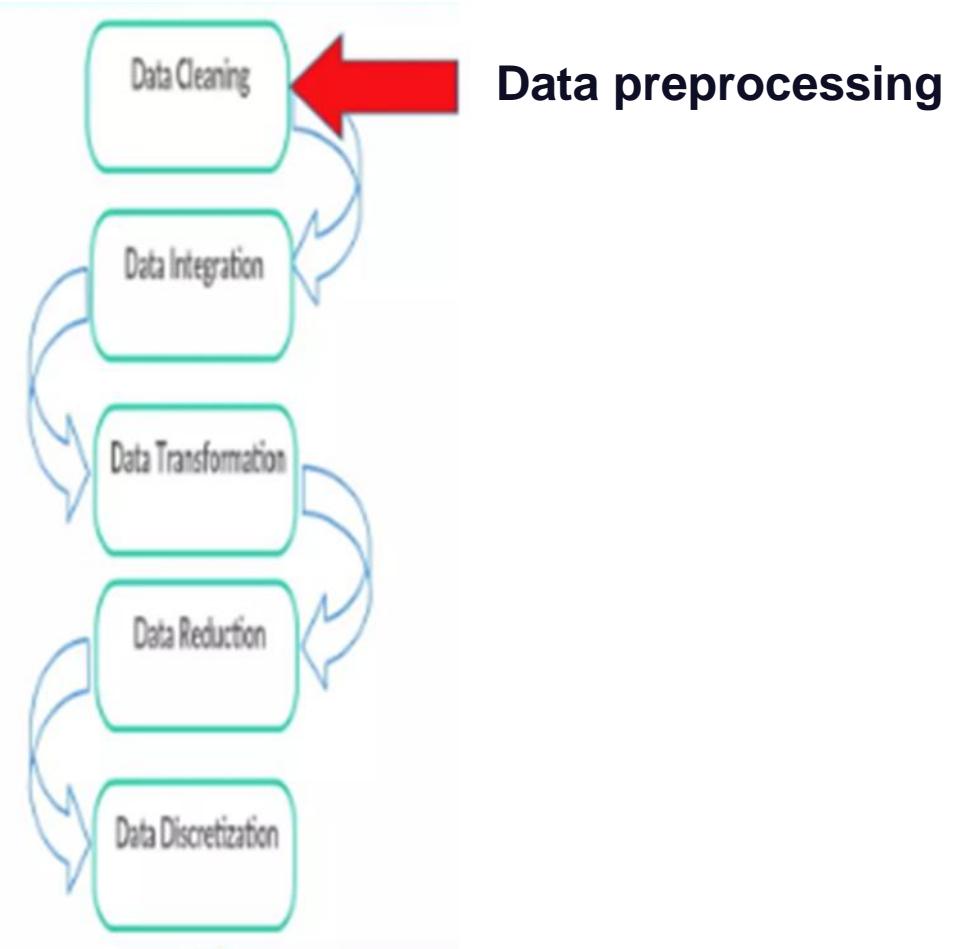
Encoding:

- Feature encoding is defined as the method to **convert string values into numeric form**.
- This is important as all ML models require all values in numeric format.
- Feature encoding includes label encoding and One Hot Encoding (also known as get_dummies).



Data Pre-Processing

1. Data cleaning
2. Data integration
3. Data transformation
4. Data reduction
5. Data Discretization





Cntd..

- Data preparation is also known as data "**pre-processing**," "**data wrangling**," "**data cleaning**," "**data pre-processing**," and "**feature engineering**."
- It is the later stage of the machine learning lifecycle, which comes after data collection..

The data preparation process can be complicated by issues such as:

1. **Missing or incomplete records:** Missing data sometimes appears as empty cells, values (e.g., NULL or N/A), or a particular character, such as a question mark

age	weight
[50-60)	?
[20-30)	[50-75)
[80-90)	?
[50-60)	?
[50-60)	?
[70-80)	?



Cntd..

- **Data is not always available**

E.g., while admission filling form by student at the time of admission, he might be don't known local guardian contact number.

- **Missing data may be due to**

- ✓ equipment malfunction
- ✓ inconsistent with other recorded data and thus deleted
- ✓ data not entered due to misunderstanding
- ✓ certain data may not be considered important at the time of entry
- ✓ no register history or changes of the data
- ✓ expansion of data schema

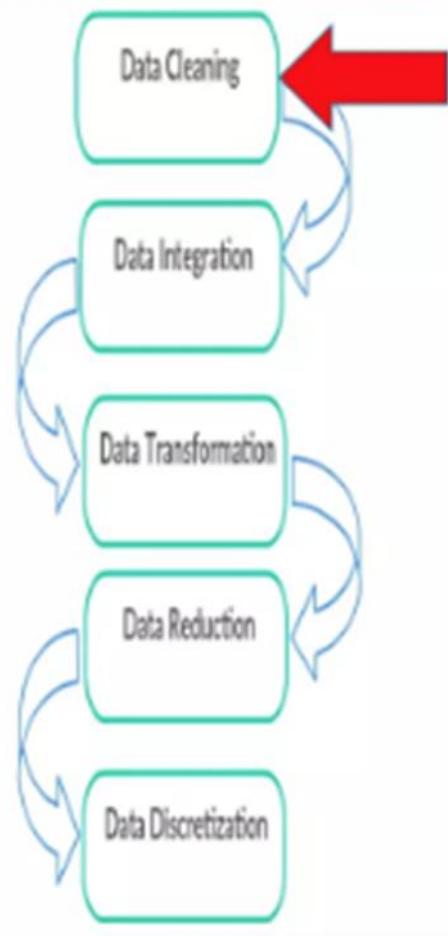


Cntd..



How to Handle Missing Data?

- Ignore the tuple (loss of information)
- Fill in missing values manually: tedious, infeasible?
- Fill in it automatically with
 - ✓ a global constant : e.g., unknown, a new class?!
 - ✓ **Imputation:** Use the attribute **mean** to fill in the missing value,
 - ✓ Use the most **probable value** to fill in the missing value.

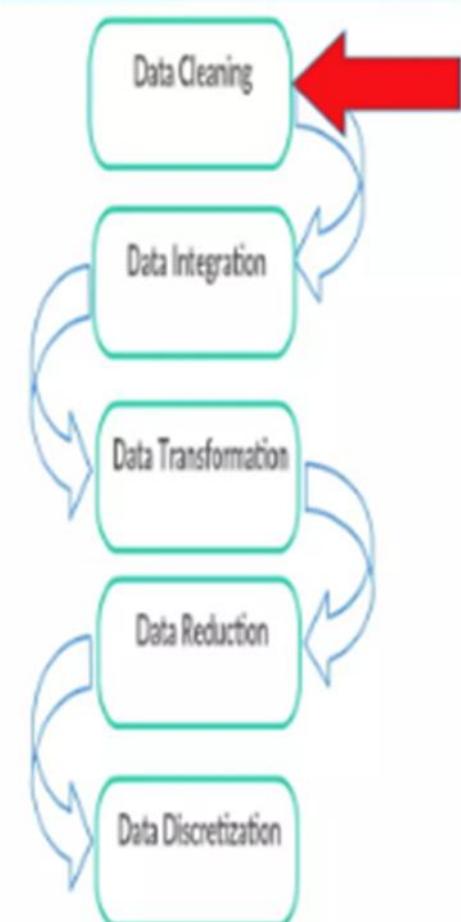


Cntd..



Noisy Data

- Noise: random **error** or variance in a measured variable
- Incorrect attribute values may due to
 - ✓ faulty data collection instruments
 - ✓ data entry problems
 - ✓ data transmission problems
 - ✓ technology limitation
 - ✓ inconsistency in naming convention
- Other data problems which requires data cleaning
 - ✓ duplicate records
 - ✓ incomplete data
 - ✓ inconsistent data

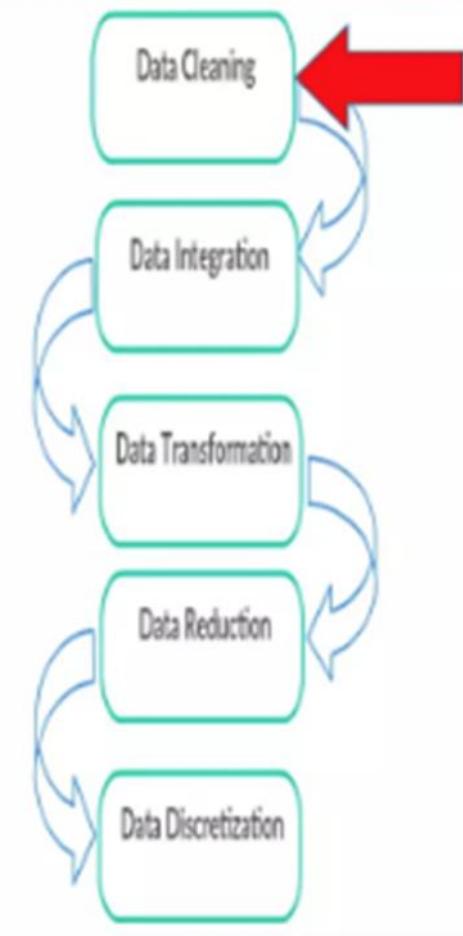




Cntd..

How to handle noisy data?

- **Binning method:**
 - ✓ first **sort data** and **partition** into (equi-depth) bins
 - ✓ then one can smooth by bin **means**, smooth by bin **median**, smooth by bin **boundaries**, etc.
- **Combined computer and human inspection**
 - ✓ detect suspicious values and check by human





Cntd..

Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into (equi-depth) bins:

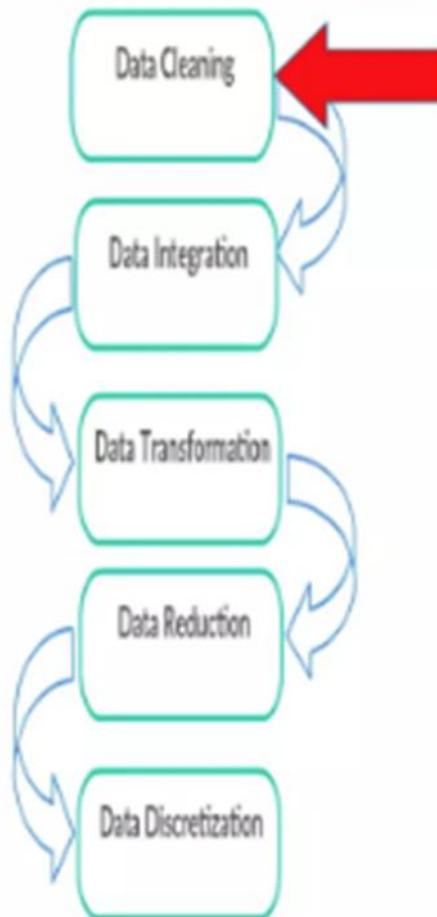
- Bin 1: **4, 8, 9, 15**
- Bin 2: **21, 21, 24, 25**
- Bin 3: **26, 28, 29, 34**

* Smoothing by bin means:

- Bin 1: **9, 9, 9, 9** $(4+8+9+15/4)=9$
- Bin 2: **23, 23, 23, 23** $(21+21+24+25/4)=23$
- Bin 3: **29, 29, 29, 29** $(26+28+29+34/4)=29$

* Smoothing by bin boundaries:

- Bin 1: **4, 4, 4, 15**
- Bin 2: **21, 21, 25, 25**
- Bin 3: **26, 26, 26, 34**



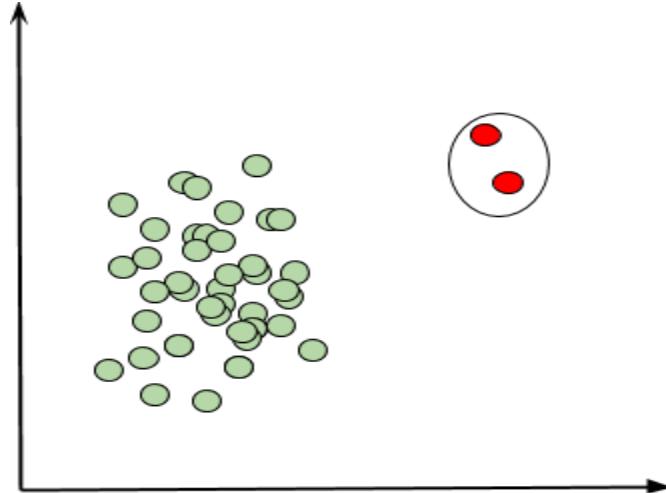


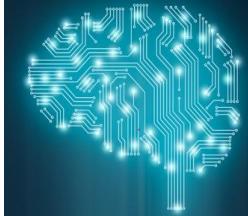
Cntd..

॥ विद्यशान्तिर्धूमं ध्रुता ॥

2. Outliers or anomalies: Unexpected values

- ML algorithms are sensitive to the range and distribution of values when data comes from unknown sources.
- These values can spoil the entire machine learning training system and the performance of the model.
- Hence, it is essential to detect these outliers or anomalies through techniques such as visualization technique.





Cntd..

3. Unstructured data format :

- Data comes from various sources and needs to be extracted into a different format.
- Hence, before deploying an ML project, always consult with domain experts or import data from known sources.

4. Limited or sparse features / attributes :

- Whenever data comes from a single source, it contains limited features,
- so it is necessary to import data from various sources for feature enrichment or build multiple features in datasets.

5. Understanding feature engineering:

- Features engineering helps develop additional content in the ML models, increasing model performance and accuracy in predictions.



Cntd..

Data Integration

Data integration:

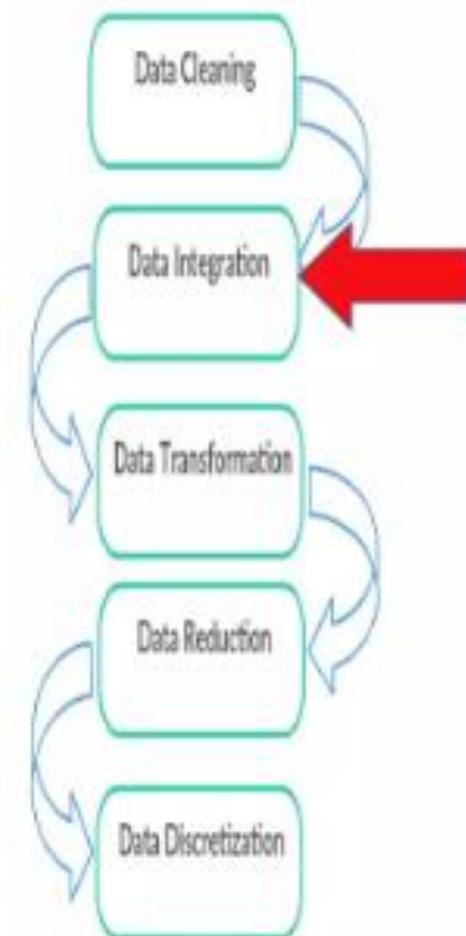
Its **combines data** from multiple sources

- **Schema integration**

Integrate metadata from different sources

Entity identification problem: identify real world entities from multiple data sources, e.g., A.cust-id \equiv B.cust-#

- **Detecting and resolving data value conflicts**
- for the same real world entity, attribute values from different sources are different, e.g., different scales, metric vs. British units
- **Removing duplicates and redundant data**



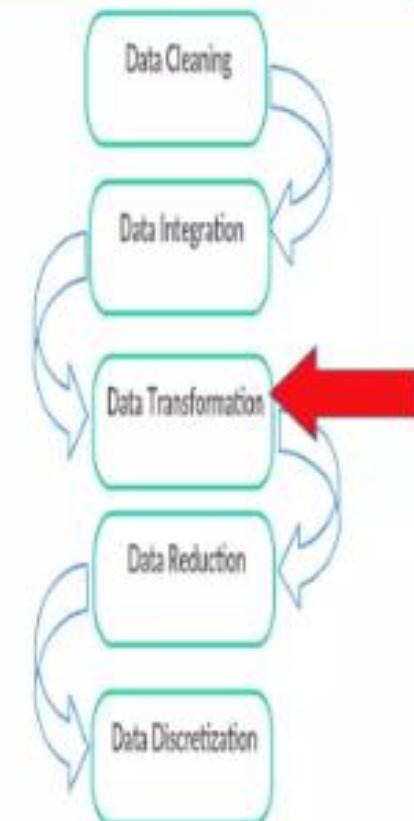


Cntd..

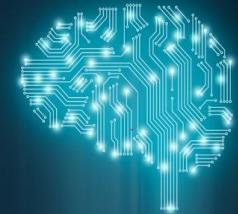
Data Transformation

Data Transformation

- Smoothing: **remove noise** from data
- **Normalization**: scaled to fall within a small, specified range
- Attribute/feature construction
 - ✓ New attributes constructed from the given ones
- **Aggregation**: summarization
 - ✓ Integrate data from different sources (tables)

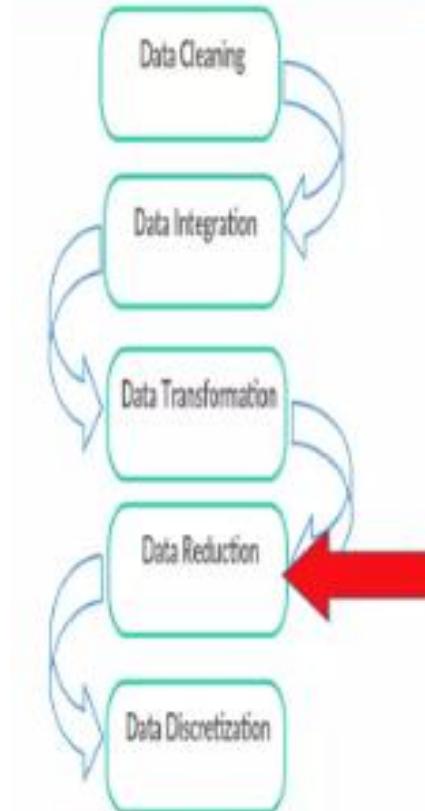


Cntd..



Data Reduction

- **Data is too big to work with**
 - ✓ Too many instances
 - ✓ too many features (attributes)
 - ✓ **Data Reduction**
 - ✓ Obtain a reduced representation of the data set that is much smaller in **volume** but yet produce the same (or almost the same) **analytical** results (easily said but difficult to do)
- **Data reduction strategies**
 - ✓ Dimensionality reduction – remove unimportant attributes
 - ✓ Aggregation and clustering -
 - ✓ Remove redundant or close associated ones
 - ✓ Sampling



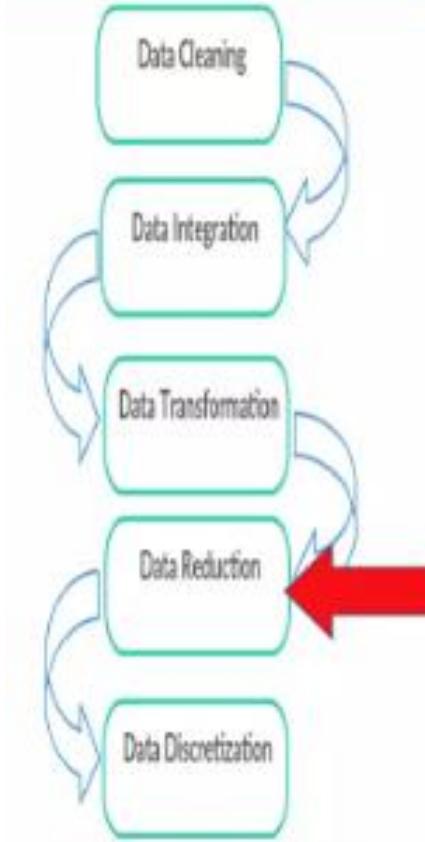


Cntd..

Data Reduction

Clustering

- Partition data set into **clusters**, and one can store cluster representation only.
- Can be very effective if data is clustered but not if data is **dirty**.
- There are many choices of clustering and clustering algorithms.



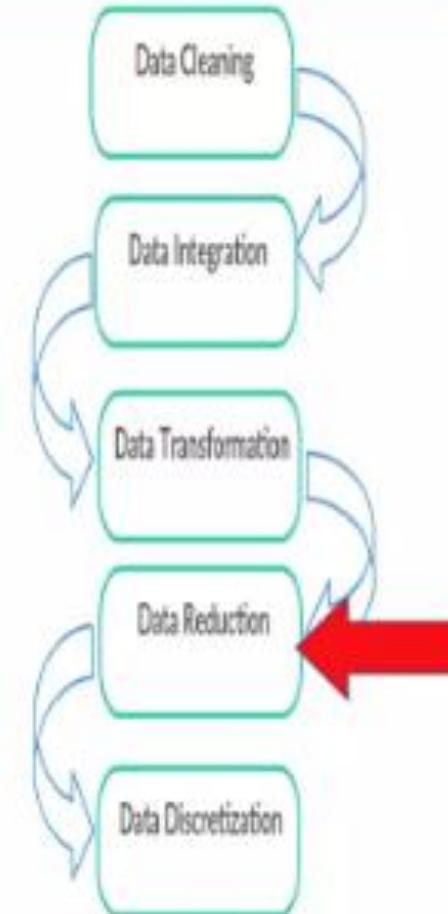


Cntd..

Data Reduction

Sampling

- **Choose a representative subset of the data**
 - ✓ Simply selecting random **sampling** may have improve performance in the presence of scenario .
- **Develop adaptive sampling methods**
 - ✓ Stratified sampling:
 - Approximate the percentage of each class (or subpopulation of interest) in the overall database

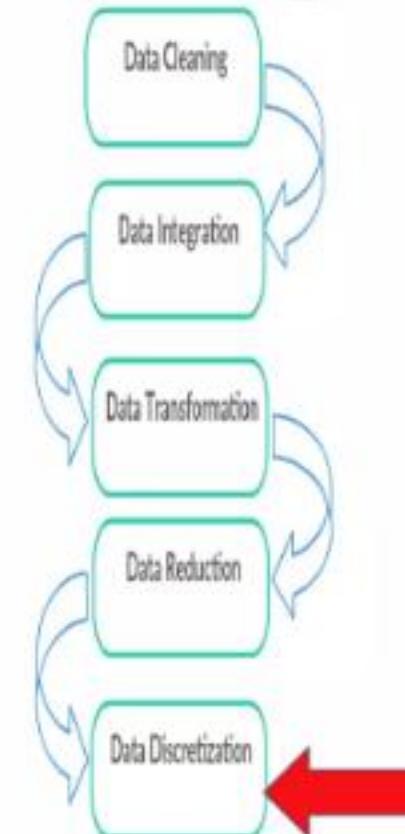




Cntd..

Data Discretization

- Discretization is a process that transforms **quantitative** data into **qualitative** data.
- It significantly improve the quality of **discovering knowledge**.
- It **reduces** the running time of various machine learning tasks such as association rule discovery, classification, clustering and prediction.
- It reduce the number of values for a given continuous attribute by dividing the range of the attribute into **intervals**.
- Interval labels can then be used to **replace** actual data values



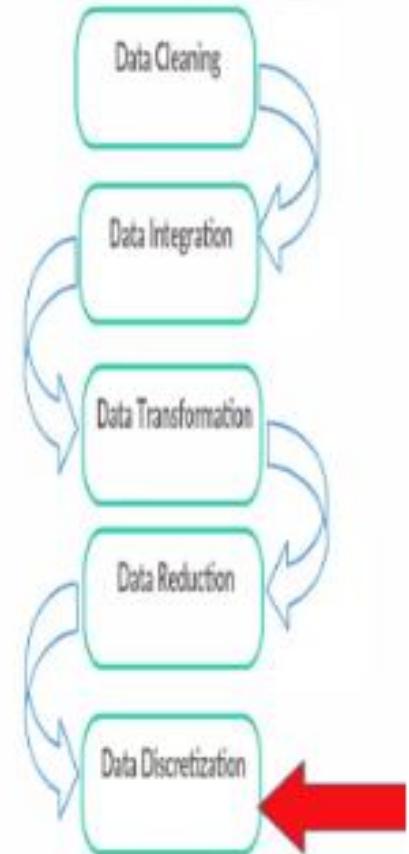
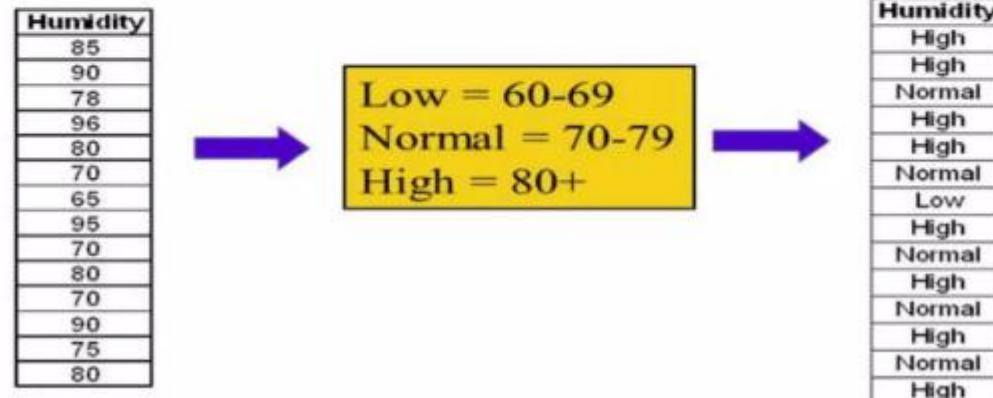


Cntd..

Data Discretization

Discretization - Example

- **Example: discretizing the “Humidity” attribute using 3 bins.**

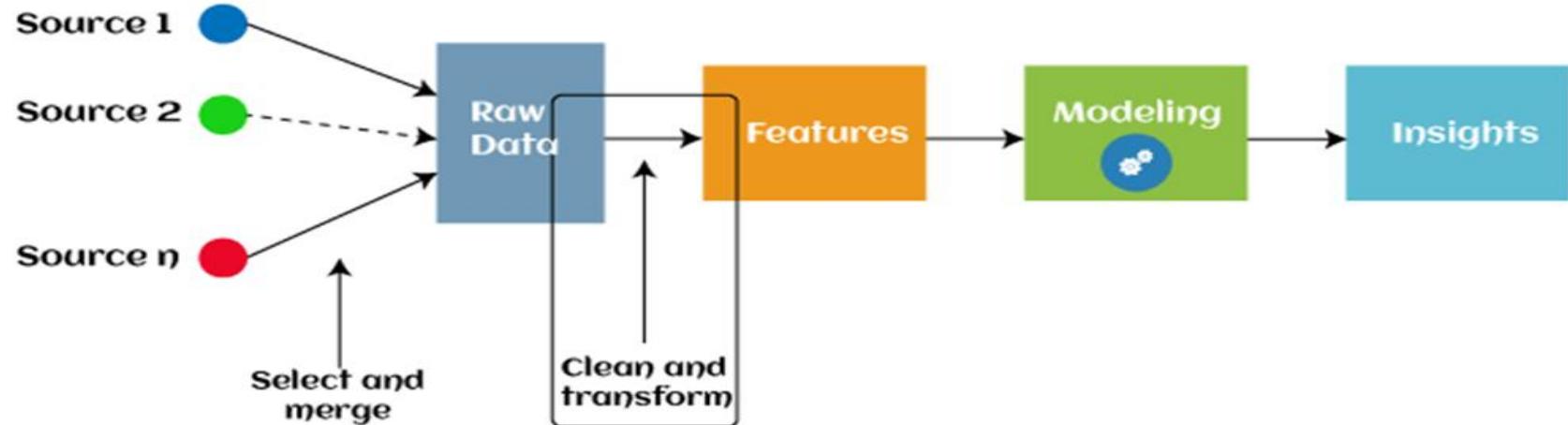




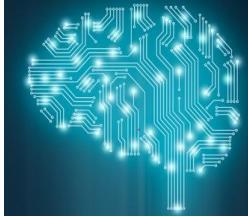
Feature Engineering

॥ विद्यानन्तर्धावं ध्रुवा ॥

Feature engineering is the pre-processing step of machine learning, which is used to **transform raw data into features** that can be used for creating a predictive model using Machine learning or statistical Modelling.



Feature engineering is the pre-processing step of machine learning, which extracts features from raw data.



Feature Engineering

What is a feature?

- Generally, all machine learning algorithms take input data to generate the output.
- The input data remains in a tabular form consisting of rows (instances or observations) and columns (variable or attributes), and these attributes are often known as **features**.

Feature Engineering processes:

1. **Feature Creation:** Feature creation is **finding the most useful variables to be used in a predictive model**.
2. **Transformations:** The transformation step of feature engineering involves **adjusting the predictor variable to improve the accuracy** and performance of the model.



Feature Engineering

3. Feature Extraction: Feature extraction is an automated feature engineering process that **generates new variables** by extracting them from the raw data.

The main aim of this step is to **reduce the volume** of data so that it can be easily used and managed for data modelling.

Feature extraction methods include **cluster analysis, text analytics, edge detection algorithms, and principal components analysis (PCA)**.

4. Feature Selection: *Feature selection is a way of selecting the subset of the most relevant features from the original features set by removing the redundant, irrelevant, or noisy features."*



Feature Engineering

Steps in Feature Engineering

Data Preparation:

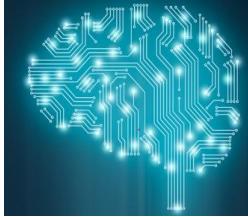
- In this step, raw data acquired from different resources are prepared **to make it in a suitable format** so that it can be used in the ML model.
- The data preparation may contain **cleaning of data, delivery, data augmentation, fusion, ingestion, or loading**.

Exploratory Analysis:

- This step involves **analysis, investing data set, and summarization** of the main characteristics of data.
- Different **data visualization techniques** are used to better understand the manipulation of data sources, to find the most appropriate statistical technique for data analysis & to select the best features for the data.

Benchmark:

- Benchmarking is a process of **setting a standard baseline** for accuracy to compare all the variables from this baseline.
- The benchmarking process is used to **improve the predictability of the model and reduce the error rate**.



Feature Engineering

Feature Engineering Techniques:

1. **Imputation:** Imputation is responsible for handling irregularities within the dataset.
 - For **numerical data imputation**, a **default value** can be imputed in a column, and missing values can be filled with **means or medians** of the columns.
 - For **categorical data imputation**, missing values can be interchanged with the maximum occurred value in a column.
2. **Handling Outliers:** This technique first identifies the outliers and then remove them out.
 - **Standard deviation** can be used to identify the outliers
 - **Z-score** can also be used to detect outliers.



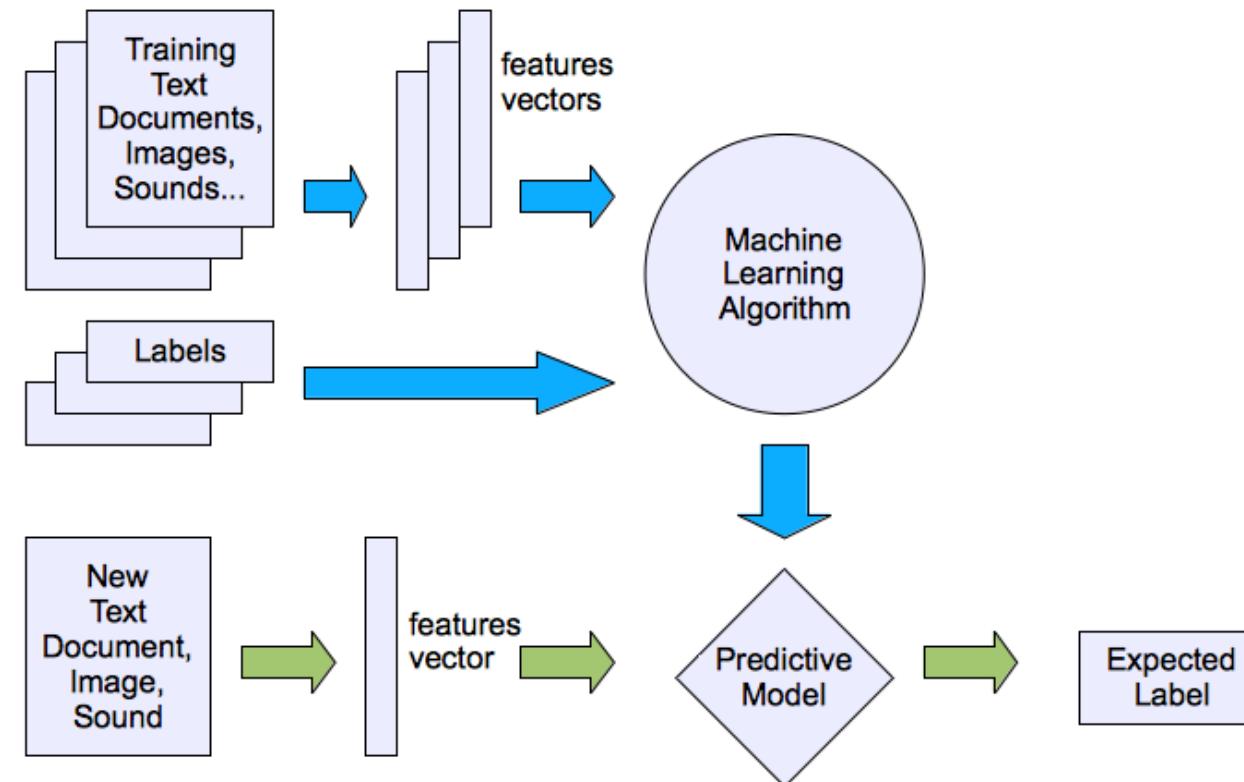
Feature Engineering

Feature Engineering Techniques:

3. **Log transform:** Log transform helps in **handling the skewed data**, and it makes the distribution more approximate to normal after transformation.
4. **Binning:** can be used to normalize the **noisy data**. This process involves segmenting different features into bins.
5. **Feature Split:** is the process of splitting features intimately into **two or more parts** and performing to make new features.
6. **One hot encoding:** It is a technique that converts the categorical data in a form so that they can be easily understood by machine learning algorithms and hence can make a good prediction.

Types of Learning

- Supervised learning





Types of Learning

Supervised Learning :

Aim is to learn a mapping from the input to an output whose correct values are provided by a supervisor.

1. **Classification** : Data is labelled ...meaning it is assigned a class, for example **spam/non-spam** or **fraud/non-fraud**

e.g. for financial institution ..**input to classifier** is **savings and income** and **output** is one the class like **high risk or low risk** based on following classification rule

- if income > δ1 and saving δ2 then low risk else high risk**

2. **Regression** : **Data is labelled** with a **real value** rather then a label.

e.g. price of a stock over time.

e.g **predict the price of used car**

Input : brand, year, engine capacity, mileage & other information ...

output: **Price of car**



Types of Learning

Types of Machine Learning

Supervised Learning



Unsupervised Learning

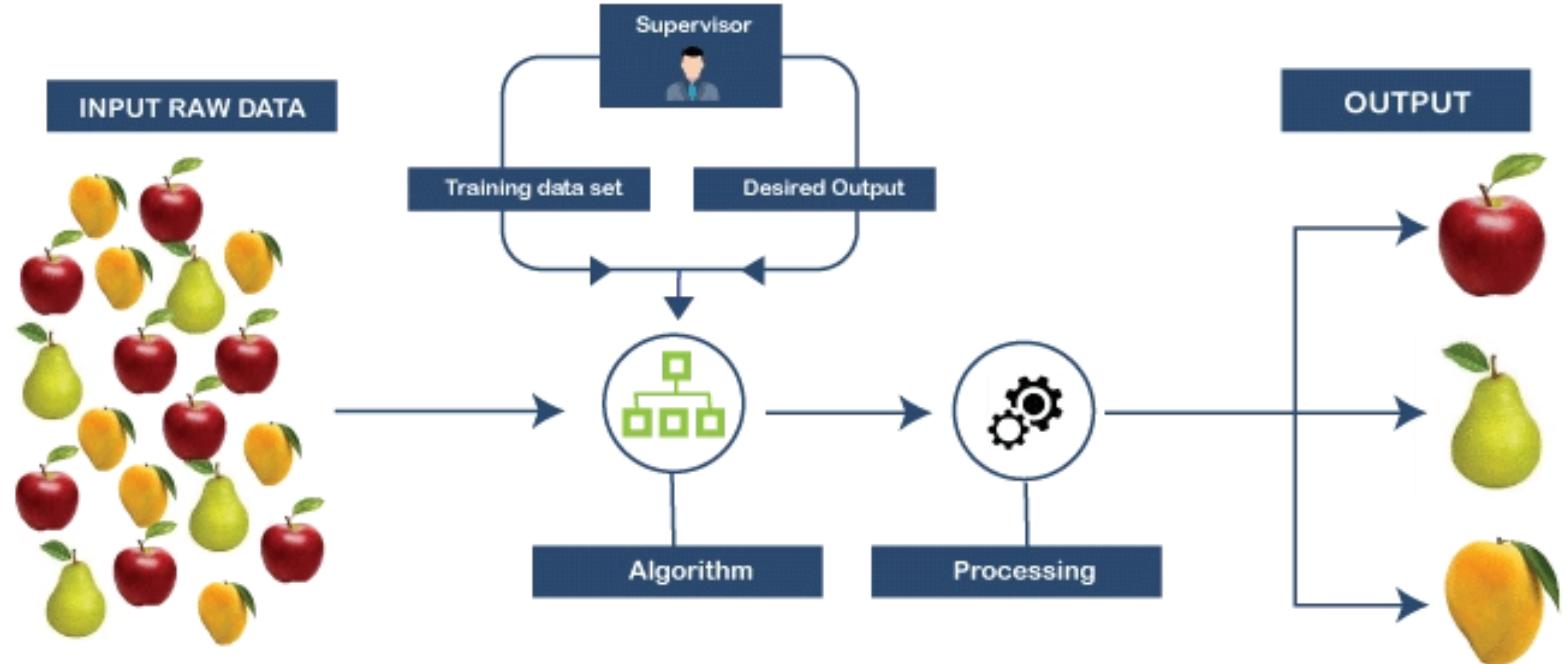


Machine Learning Algorithms



Types of Learning

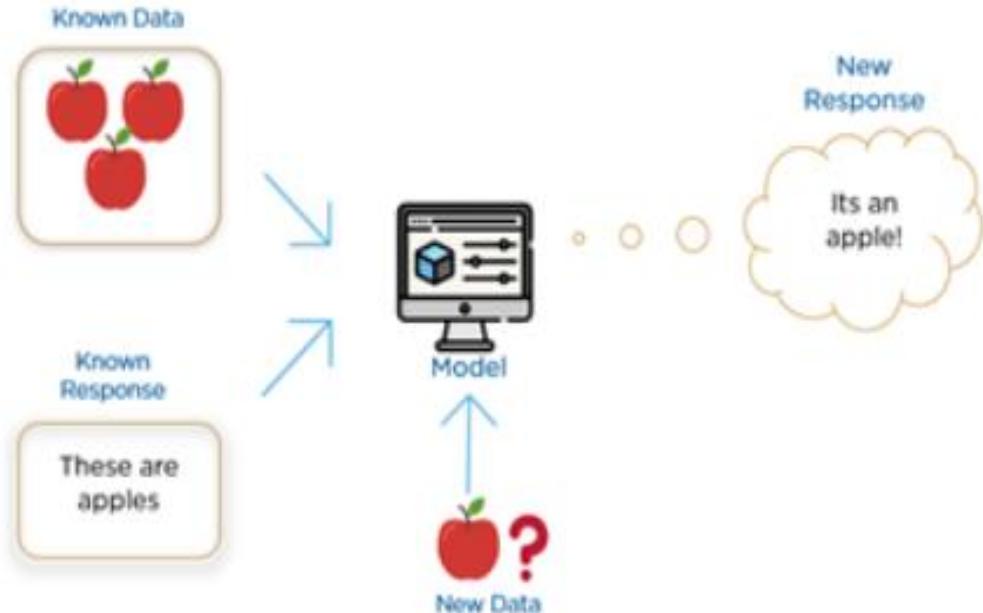
SUPERVISED LEARNING





Types of Learning

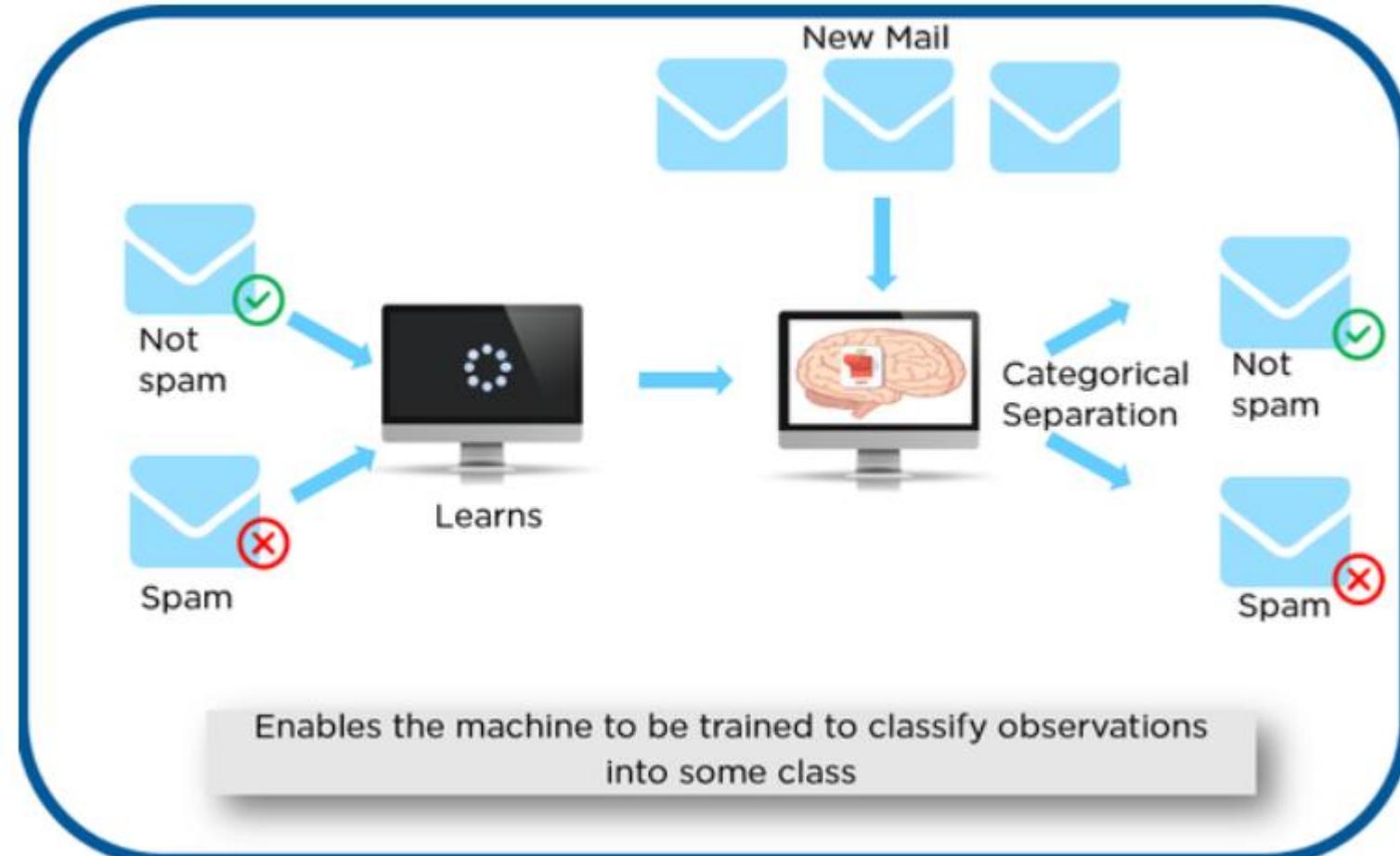
Supervised Learning :





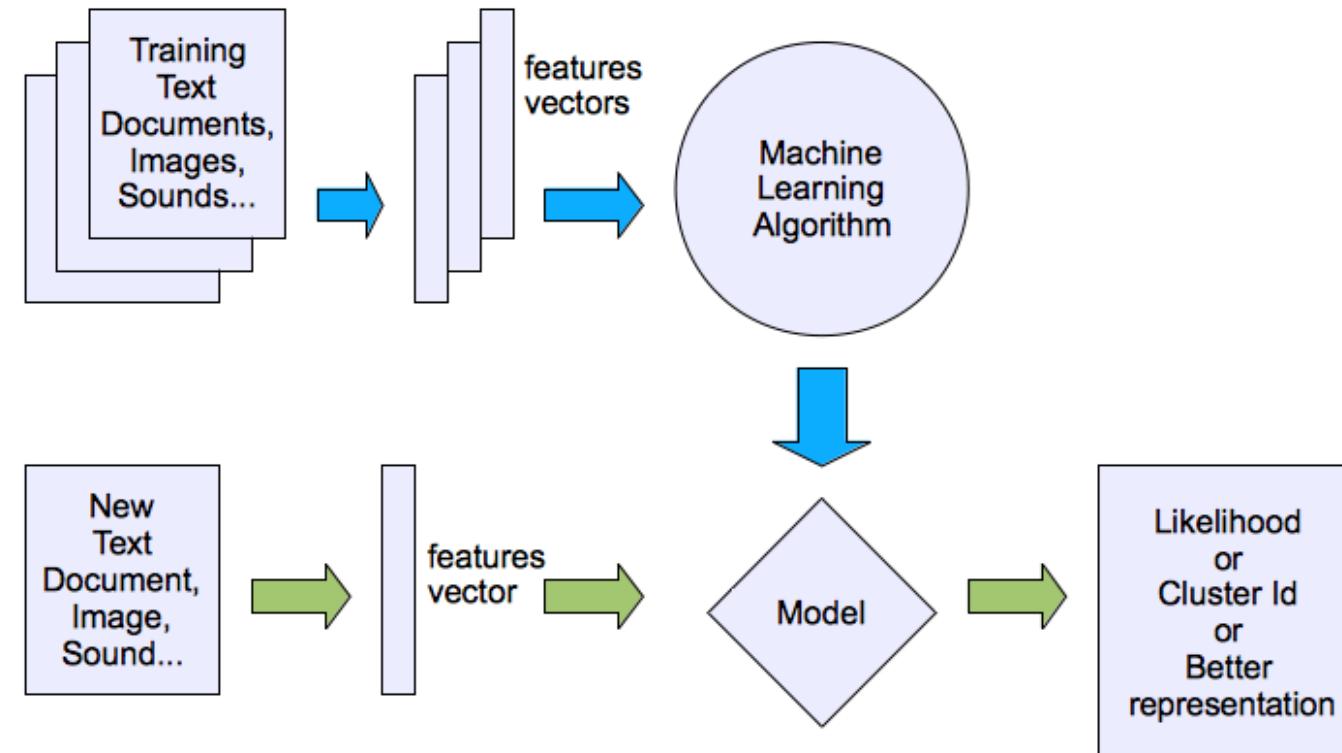
Types of Learning

Supervised Learning :



Unsupervised Learning

- Unsupervised learning





MIT-WPU

॥ विद्यशान्तिर्धूमं धूता ॥



Example of Unsupervised learning

- Clustering
- Association

Clustering



sample

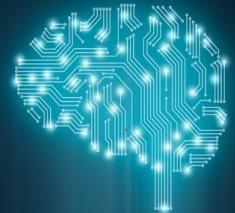


Cluster/group



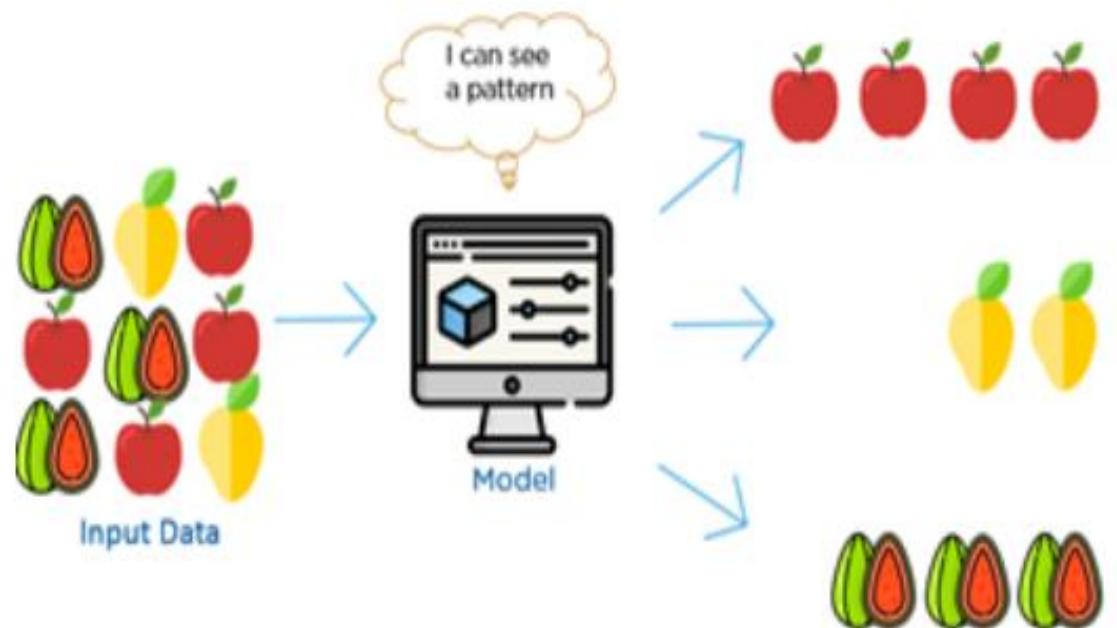
MIT-WPU

॥ विद्यशान्तिर्धूवं ध्रुवा ॥



Example of Unsupervised learning

- Clustering
- Association

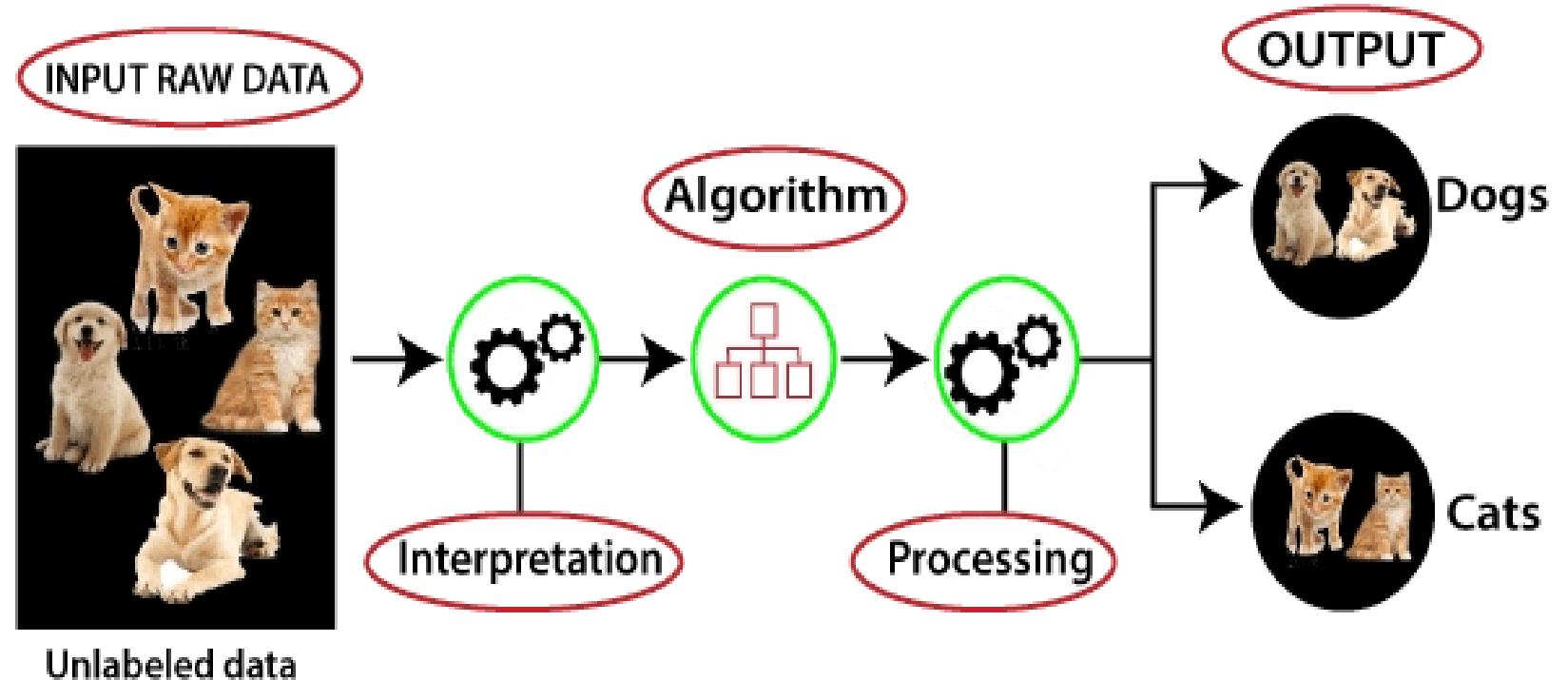




MIT-WPU
॥ विद्यशान्तिर्धूमं धूवा ॥



Example of Unsupervised learning

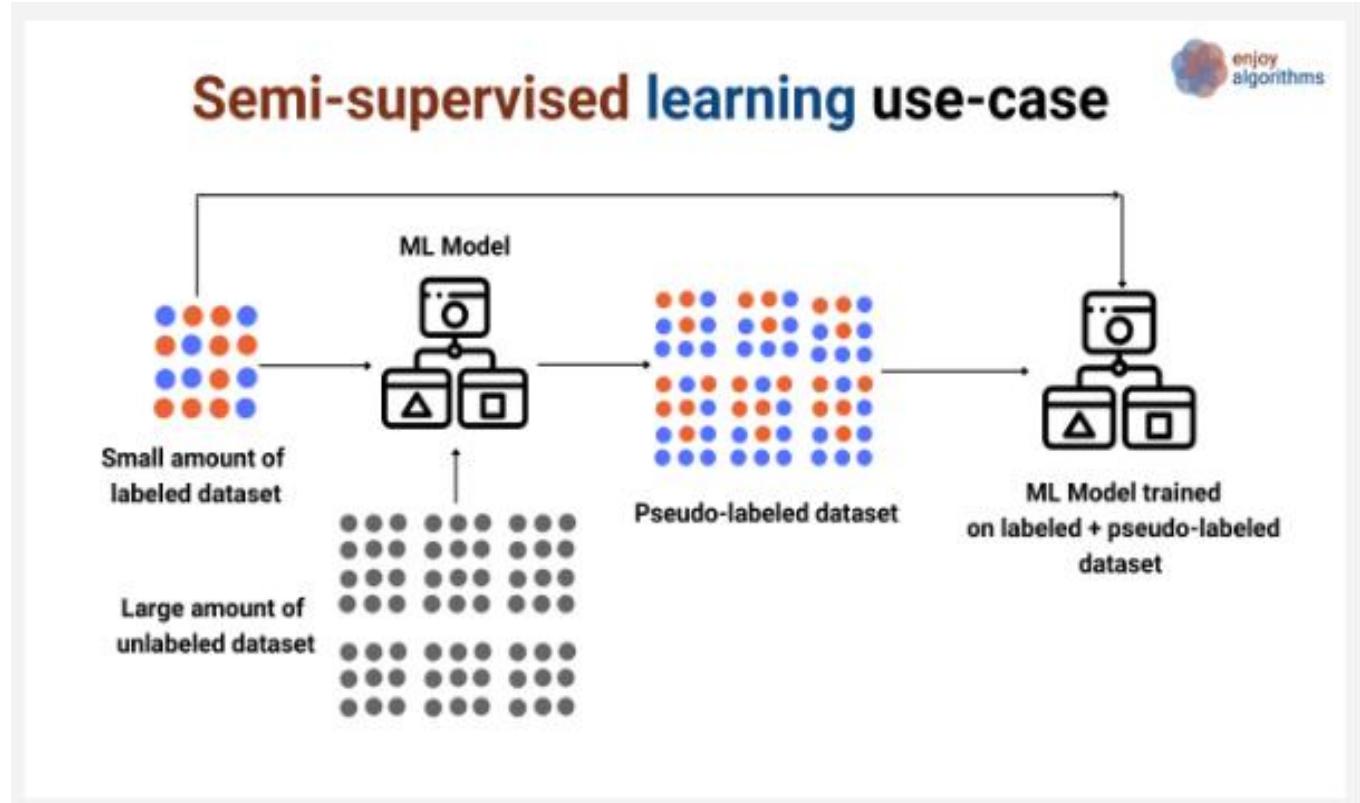




MIT-WPU
॥ विद्यशान्तिर्धूमं ध्रुवा ॥



Example of Semi-supervised learning



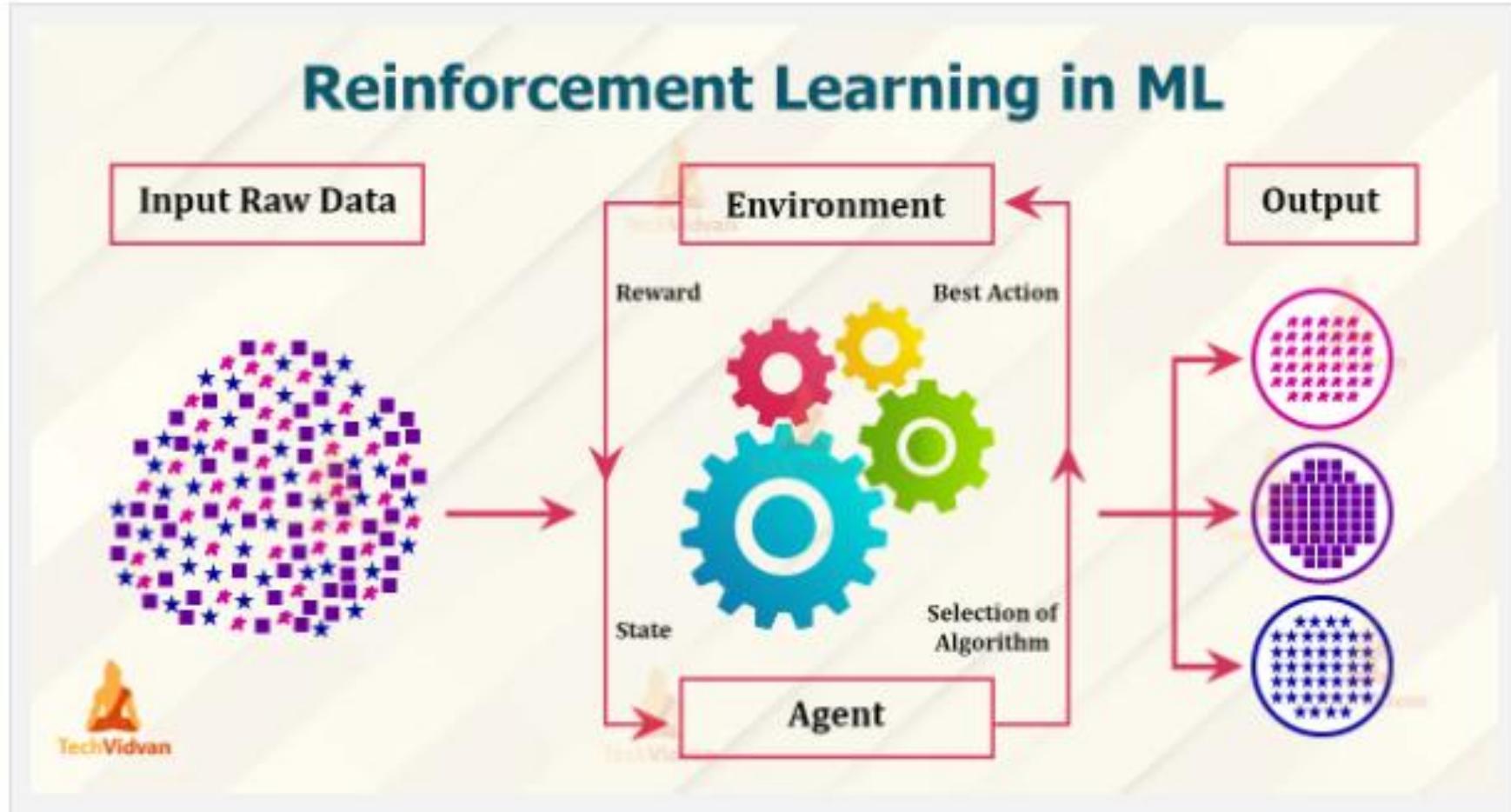


Reinforcement Learning

- learning from mistakes
- Place a reinforcement learning algorithm into any environment and it will make a **lot of mistakes in the beginning**
- As we provide some sort of signal to the algorithm that associates **good behaviors with a positive signal** and **bad behaviors with a negative one**
- we can **reinforce our algorithm to prefer good behaviors over bad ones.**
- Over time, our learning algorithm learns to make less mistakes than it used to.



Reinforcement Learning

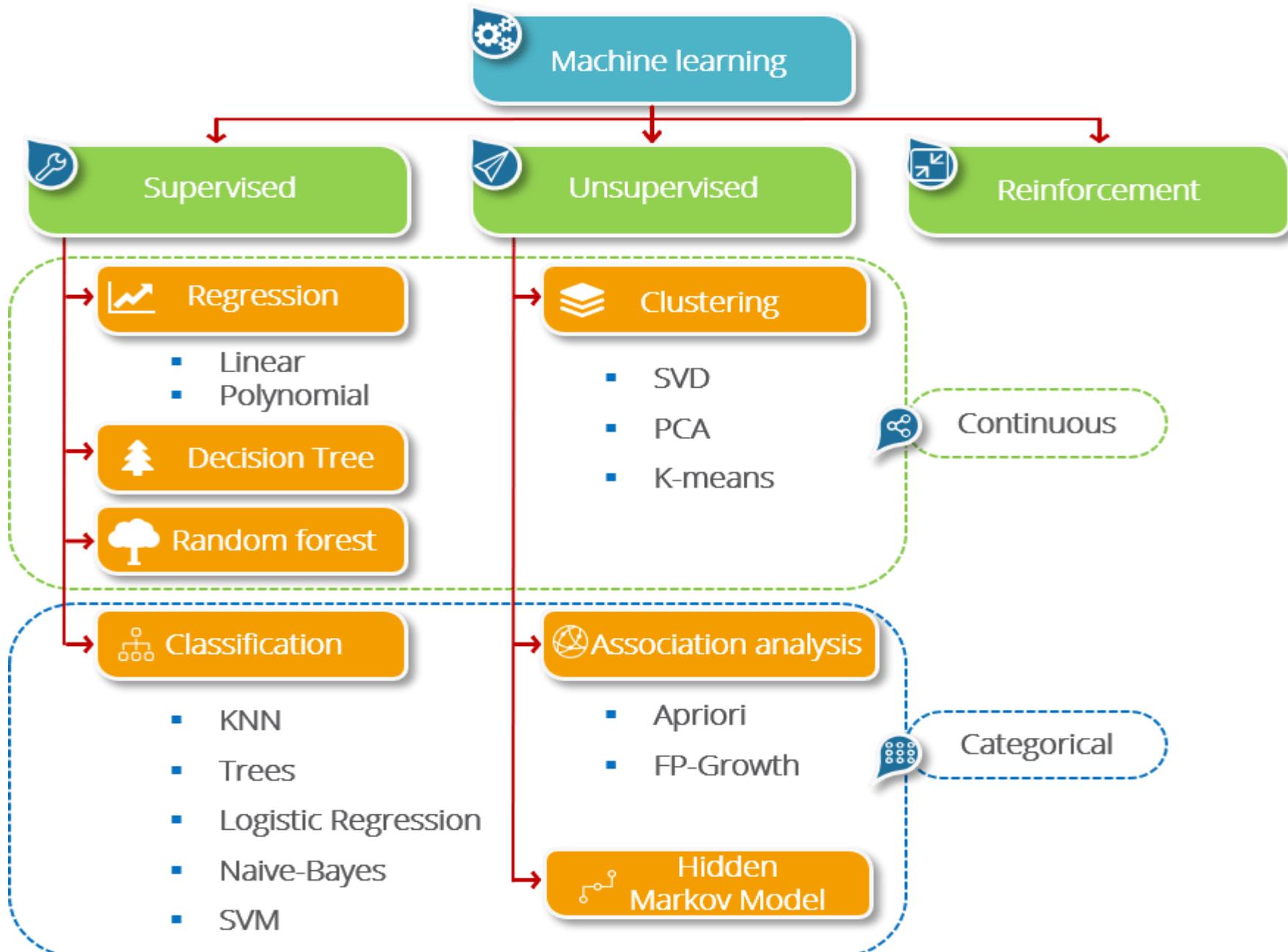




Reinforcement Learning

Where is reinforcement learning in the real world?

- **Video Games**
- **Industrial Simulation:**
- **Resource Management**





Key Elements of Machine Learning

- There are tens of thousands of machine learning algorithms and hundreds of new algorithms are developed every year.
- Every machine learning algorithm has three components:

1. Representation: how to represent knowledge.

Examples include decision trees, sets of rules, instances, graphical models, neural networks, support vector machines, model ensembles and others.

2. Evaluation: the way to evaluate candidate programs (hypotheses).

Examples include accuracy, prediction and recall, squared error, likelihood, posterior probability, cost, margin, entropy k-L divergence and others.



3. Optimization: the way candidate programs are generated **known as the search process.**

For example combinatorial optimization, convex optimization, constrained optimization.

- All machine learning algorithms are combinations of these three components.
- A framework for understanding all algorithms.



Aspects of developing a learning system:

training data, concept representation, function approximation

- For training and testing purpose of our model we need to split the dataset in to three distinct dataset, **training set, validation set and testing set**
- **Training set:-**
- A set of data used **to train the model**
- It is used **to fit the model**
- The model sees and **learn from this data**
- **Later on the trained model can be deployed and used to accurately predict on new data that it has not seen before**
- **Labeled data is used**



Validation set

- Validation set is the set of **data separate from the training data**
- It is used to **validate our model during training**
- It **gives information** which is used for tuning model hyper parameter
- It **ensures that our model is not over fitting** to the data in the training set
- Labeled data is used



Test Set

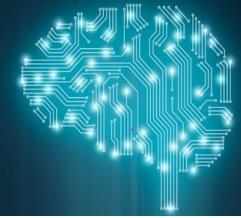
- A set of data used to **test the model**
- The test set is **separated from both the train set and validation set**
- Once the model is trained and validated using the training data and validation sets then the **model is used to predict the output** for the data in the test set
- **Unlabeled data is used**



Data Split

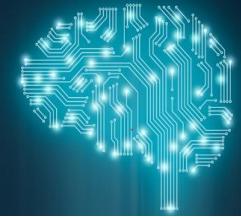


- Rules for performing data split operation
- In order to avoid a correlation between the original dataset must be **randomly shuffled** before applying the split phase
- All the split must represent the original distribution
- The percentage of splitting is mostly **60%** for training **20%** for validation and **20%** for testing
- With scikit-learn this can be done using **train_test_split()** function



Exploratory Data Analysis

Exploratory Data Analysis refers to the **critical process of performing initial investigations on data** so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions **with the help of summary statistics and graphical representations.**



Exploratory Data Analysis

Typical graphical techniques used in EDA are:

- Box plot
- Histogram
- Multi-vari chart
- Run chart
- Pareto chart
- Scatter plot
- Stem-and-leaf plot
- Stem-and-leaf plot
- Parallel coordinates
- Odds ratio
- Targeted projection pursuit



EDA Example

- Wine quality data set from UCI ML repository
- imported necessary libraries (for this example pandas, numpy, matplotlib and seaborn) and loaded the data set.

```
In [2]: df = pd.read_csv('winequality-white.csv',sep=';')
df.head()
```

Out[2]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
1	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
2	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
3	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
4	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6

EDA



- Original data is separated by delimiter “ ; ” in given data set.
- To take a closer look at the data took help of “ **.head()**”function of pandas library which returns first five observations of the data set. Similarly “**.tail()**” returns last five observations of the data set.

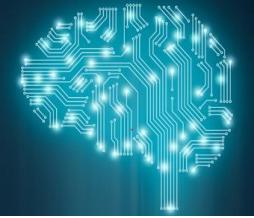


EDA Techniques

```
In [3]: df.shape
```

```
Out[3]: (4898, 12)
```

- found out the total number of rows and columns in the data set using “.shape”
- Dataset comprises of 4898 observations and 12 characteristics.
- Out of which one is dependent variable and rest 11 are independent variables — physico-chemical characteristics.
- It is also a good practice to know the columns and their corresponding data types, along with finding whether they contain null values or not.



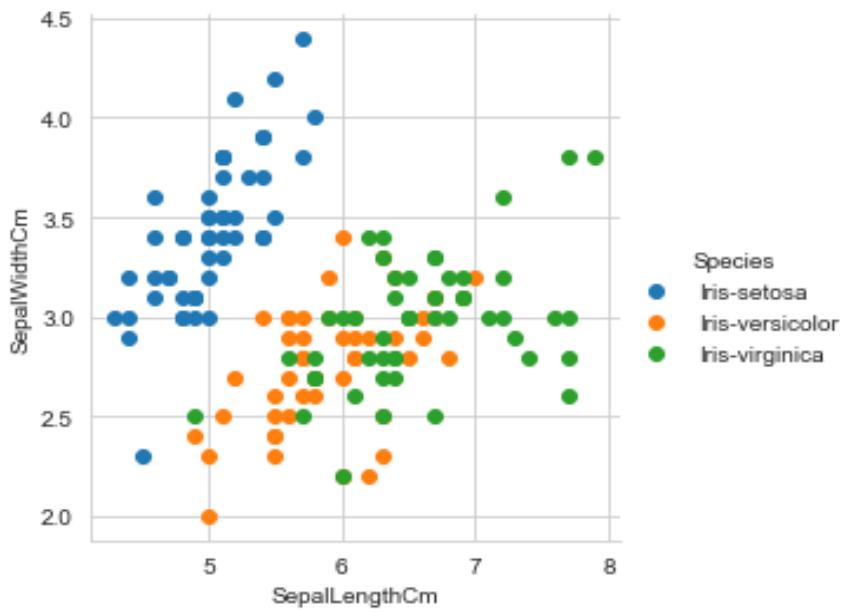
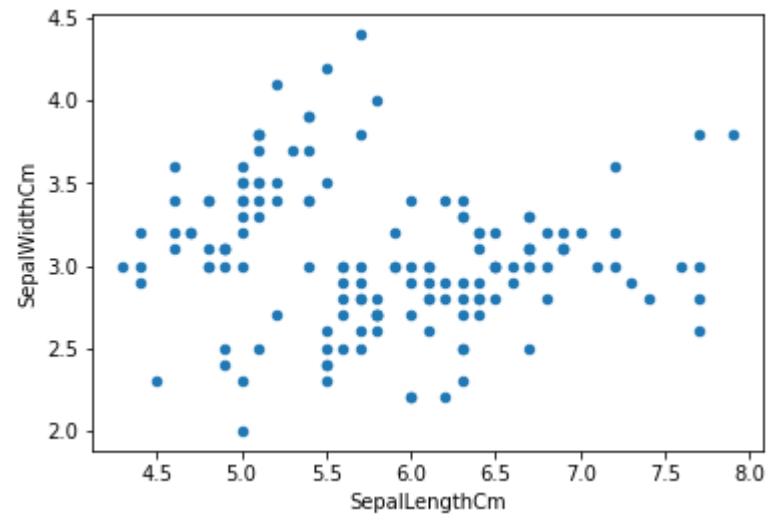
EDA: Exploratory Data Analysis

Plotting using matplotlib

```
import pandas as pd  
  
import matplotlib.pyplot as plt  
  
iris = pd.read_csv("iris.csv")  
  
iris.head(5)  
  
iris.plot(kind='scatter', x='sepal_length',  
y='sepal_width') ;  
  
plt.show()
```

import seaborn as sns

```
# 2-D Scatter plot with color-coding for each flower  
type/class.  
# Here 'sns' corresponds to seaborn.  
sns.set_style("whitegrid");  
sns.FacetGrid(iris, hue="species", size=4) \  
.map(plt.scatter, "sepal_length", "sepal_width") \  
.add_legend();  
plt.show();
```





EDA: Exploratory Data Analysis

3D scatter plot <https://plot.ly/pandas/3d-scatter-plots/>

```
import plotly  
import plotly.express as px  
iris = px.data.iris()  
  
fig = px.scatter_3d(iris, x='sepal_length',  
                    y='sepal_width', z='petal_width',  
                    color='species')  
  
fig.show()
```

What about 4-D, 5-D or n-D scatter plot?

Pair-plot

#Only possible to view 2D patterns.

```
plt.close();
```

```
sns.set_style("whitegrid");
```

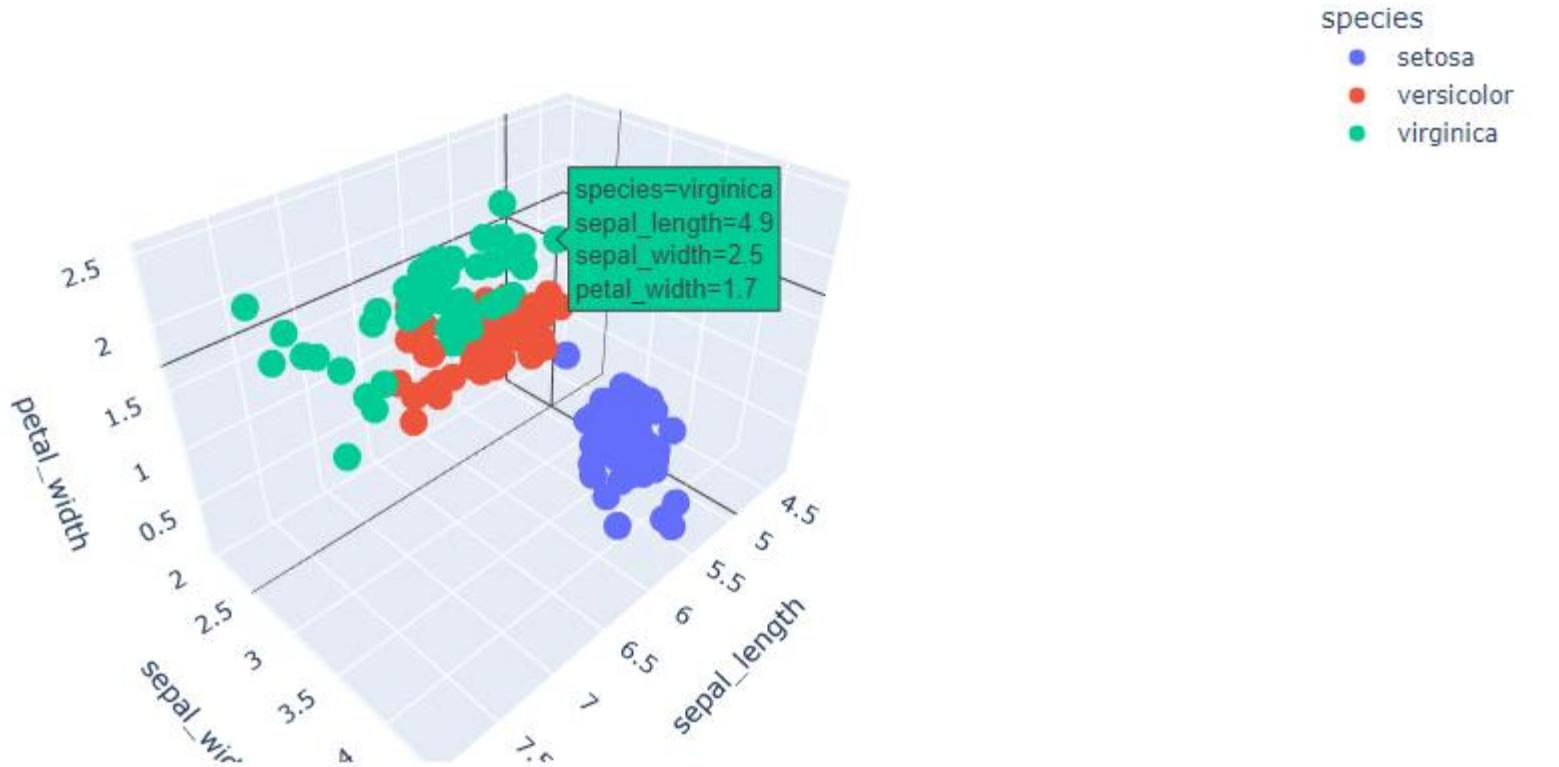
```
sns.pairplot(iris, hue="species", size=3);
```

```
plt.show()
```

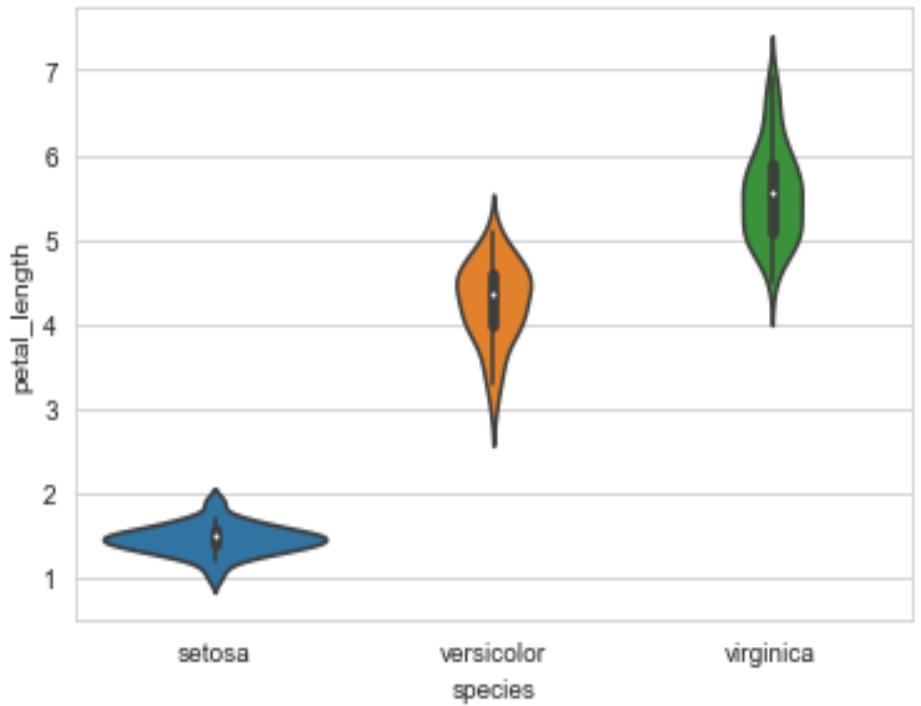
VIOLIN PLOT

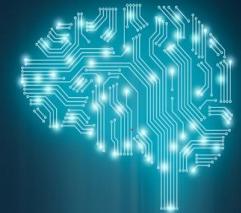
```
sns.violinplot(x="species", y="petal_length", data=iris,  
size=8)
```

```
plt.show()
```









Progressive Data Analysis

- Data has only **float** and **integer values**.
- **No** variable column has **null/missing values**.

```
In [5]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4898 entries, 0 to 4897
Data columns (total 12 columns):
fixed acidity           4898 non-null float64
volatile acidity        4898 non-null float64
citric acid              4898 non-null float64
residual sugar           4898 non-null float64
chlorides                4898 non-null float64
free sulfur dioxide      4898 non-null float64
total sulfur dioxide     4898 non-null float64
density                  4898 non-null float64
pH                       4898 non-null float64
sulphates                 4898 non-null float64
alcohol                   4898 non-null float64
quality                   4898 non-null int64
dtypes: float64(11), int64(1)
memory usage: 459.3 KB
```



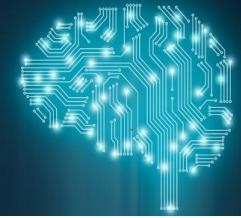
PDA Techniques

In [6]: df.describe()

Out[6]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
count	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000
mean	6.854788	0.278241	0.334192	6.391415	0.045772	35.308085	138.360657	0.994027	3.188267	0.489847	10.514267
std	0.843868	0.100795	0.121020	5.072058	0.021848	17.007137	42.498065	0.002991	0.151001	0.114126	1.230621
min	3.800000	0.080000	0.000000	0.600000	0.009000	2.000000	9.000000	0.987110	2.720000	0.220000	8.000000
25%	6.300000	0.210000	0.270000	1.700000	0.036000	23.000000	108.000000	0.991723	3.090000	0.410000	9.500000
50%	6.800000	0.260000	0.320000	5.200000	0.043000	34.000000	134.000000	0.993740	3.180000	0.470000	10.400000
75%	7.300000	0.320000	0.390000	9.900000	0.050000	46.000000	167.000000	0.996100	3.280000	0.550000	11.400000
max	14.200000	1.100000	1.660000	65.800000	0.346000	289.000000	440.000000	1.038980	3.820000	1.080000	14.200000

- The describe() function in pandas is very handy in getting various summary statistics.
- This function returns the **count, mean, standard deviation, minimum and maximum values and the quantiles** of the data.



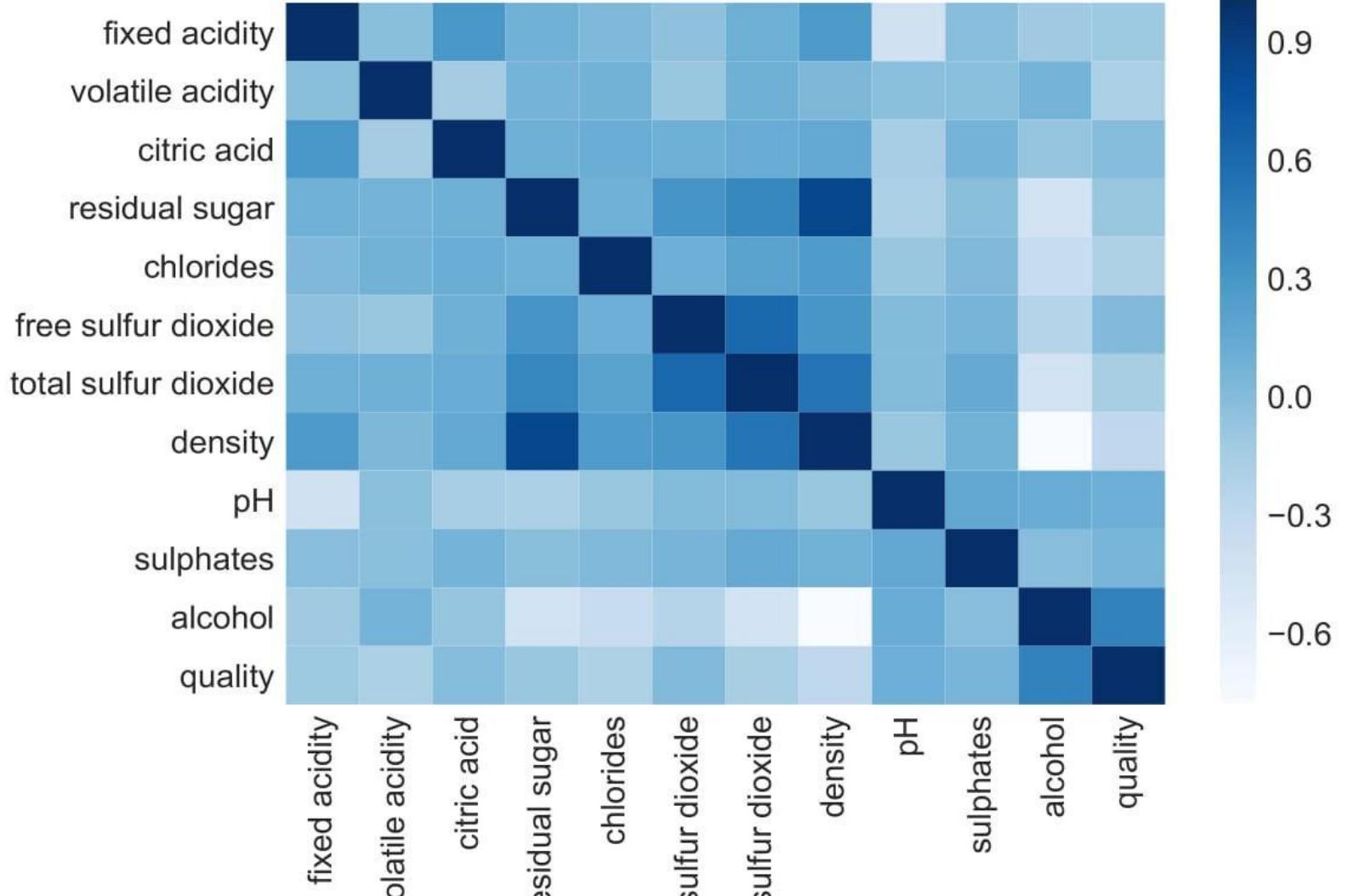
Data Preparation: Types of Data

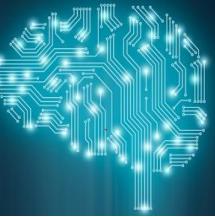
- Here as you can notice mean value is larger than median value of each column which is represented by 50%(50th percentile) in index column.
- There is notably a large difference between 75th %tile and max values of predictors “residualsugar”, ”freesulfurdioxide”, ”totalsulfur dioxide”.
- Thus observations 1 and 2 suggests that there are extreme values- Outliers in our data set.



Graph Visualisation Techniques

- Let's now explore data with beautiful graphs. Python has a visualization library , **Seaborn** which build on top of **matplotlib**.
- It provides very attractive statistical graphs in order to perform both **Univariate** and **Multivariate analysis**.
- To use linear regression for modelling, Its **necessary to remove correlated variables to improve your model**.
- One can find correlations using pandas “**.corr()**” function and can visualize the correlation matrix using a **heatmap in seaborn**.





॥ विद्यशान्तिर्धूमं ध्रुवा ॥

- Dark shades represents **positive correlation** while lighter shades represents **negative correlation**.
- If you set annot=True, you'll get values by which features are correlated to each other in grid-cells.



MIT-WPU

॥ विद्यशान्तिर्धूमं ध्रुवा ॥





Box Plot

- A box plot (or box-and-whisker plot) shows **the distribution of quantitative data** in a way that facilitates comparisons between variables.
- The **box shows the quartiles of the dataset** while the **whiskers extend to show the rest of the distribution**.
- The box plot (a.k.a. box and whisker diagram) is a standardized way of displaying the distribution of data based on the five number summary



MIT-WPU

॥ विद्यशान्तिर्धूवं ध्रुवा ॥

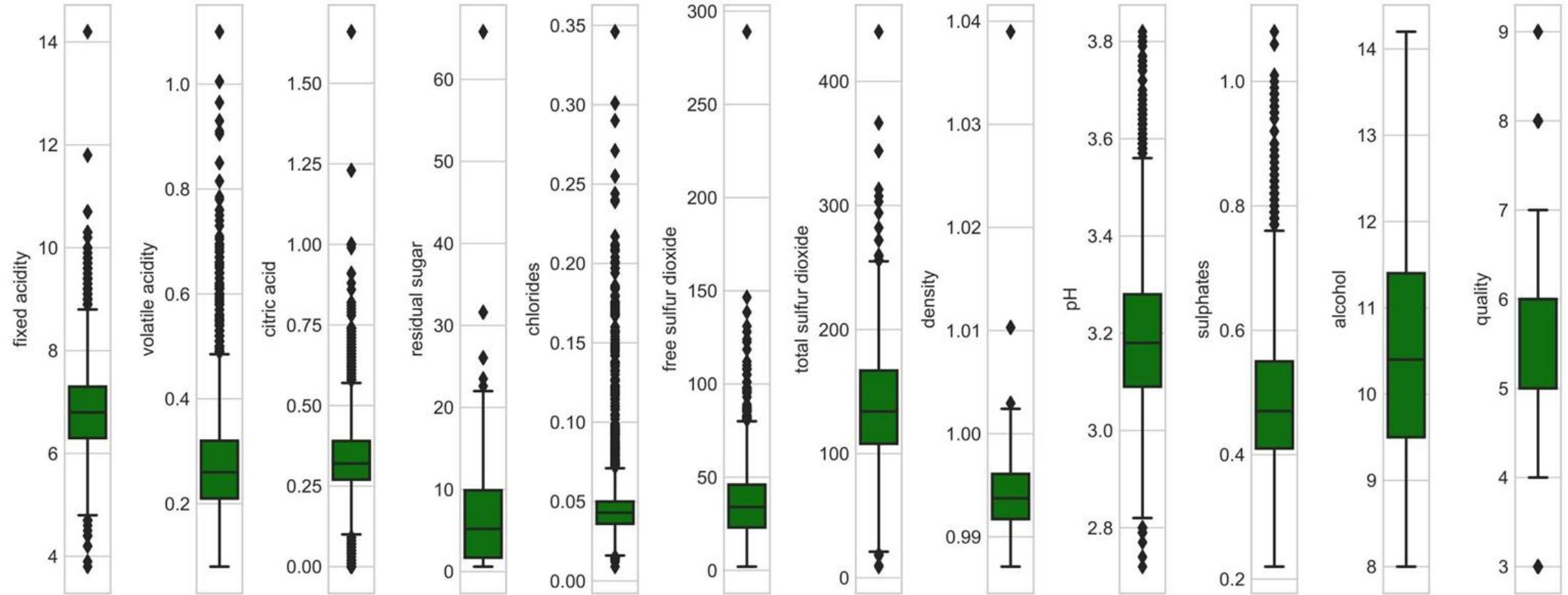


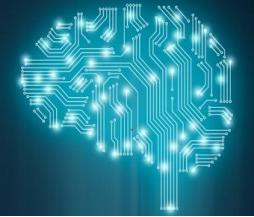
- Minimum
- First quartile
- Median
- Third quartile
- Maximum.
- In the simplest box plot the central rectangle spans the first quartile to the third quartile (the interquartile range or IQR).



MIT-WPU

॥ विद्यानन्तर्धावं ध्रुवा ॥

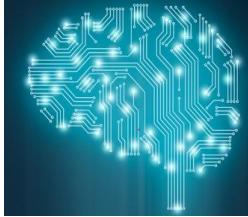




Sparse Matrix

- In numerical analysis and scientific computing, a **sparse matrix** or **sparse array** is a matrix **in which most of the elements are zero**.
- By contrast, if most of the elements are nonzero, then the matrix is considered **dense**.
- The number of zero-valued elements divided by the total number of elements (e.g., $m \times n$ for an $m \times n$ matrix) is called the **sparsity** of the matrix (which is equal to 1 minus the **density** of the matrix).
- Using those definitions, a matrix will be sparse when its Sparsity is greater than 0.5.
- Conceptually, Sparsity corresponds to systems with few pairwise interactions

Feature Engineering: Feature selection



- In machine learning and statistics, **feature selection**, also known as **variable selection**, **attribute selection** or **variable subset selection**,
- It is the **process of selecting a subset of relevant features (variables, predictors)** for use **in model construction**.
- **Feature selection** is the process of **reducing the number of input variables** when developing a predictive model.
- It is desirable to **reduce the number of input variables** to both **reduce the computational cost of modeling** and, in some cases, to **improve the performance of the model**.



Feature selection is **primarily focused** on **removing non-informative or redundant predictors** from the model.

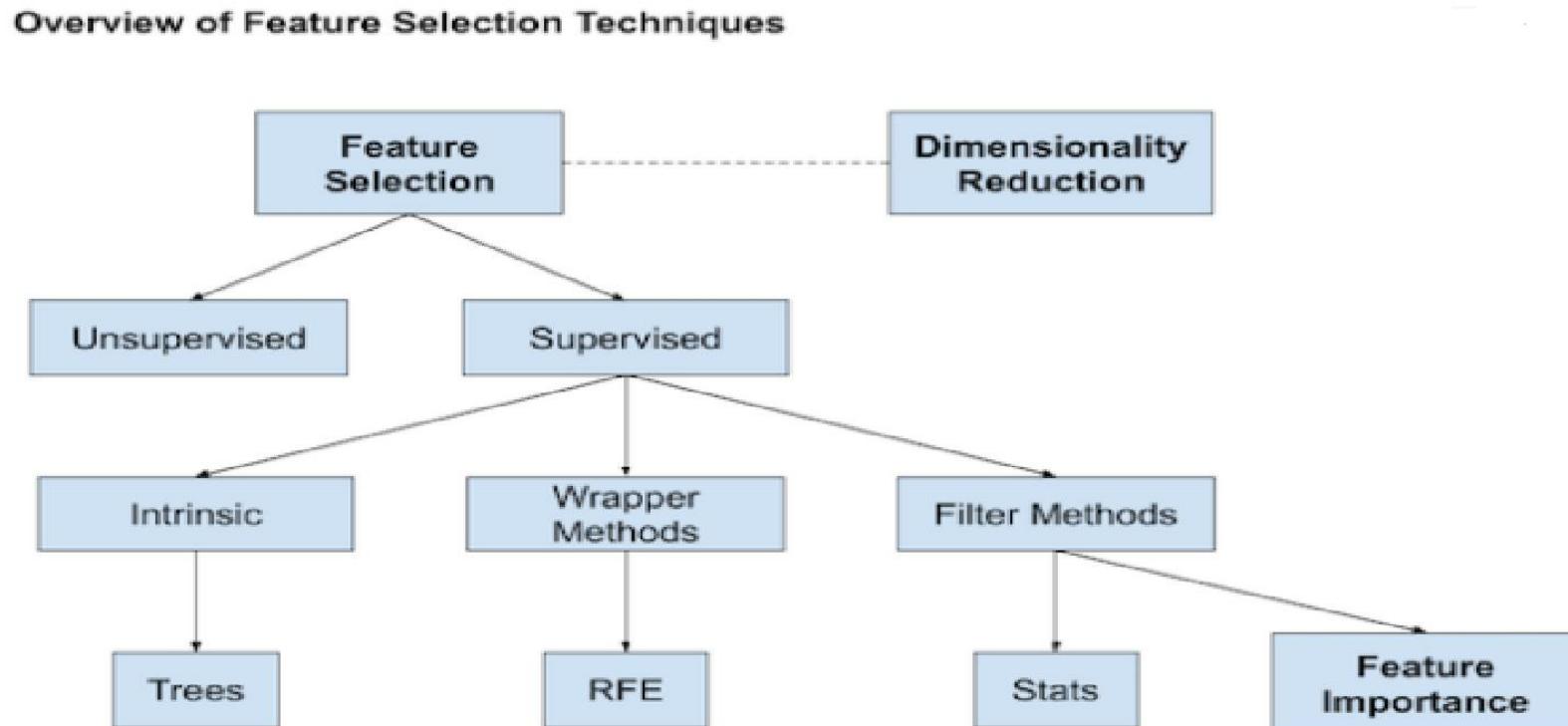
Feature selection techniques are used for several reasons:

- **simplification of models** to make them easier to interpret by researchers/users
- **shorter training times**
- to avoid the **curse of dimensionality**
- **enhanced generalization** by reducing overfitting



Feature Engineering: Feature selection

- There are **two main types of feature selection** techniques: **supervised and unsupervised**
- **Supervised methods** may be divided into **wrapper, filter and intrinsic**



Feature Engineering: Feature selection



- **Unsupervised feature selection** techniques **ignores the target variable**, such as methods that remove redundant variables using **correlation**.
- **Supervised feature selection** techniques **use the target variable**, such as methods that **remove irrelevant variables**.

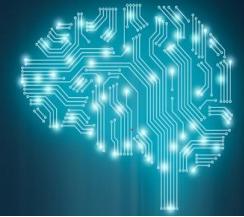
Supervised Feature Selection Techniques:

- **Filter:** Select subsets of features based on their relationship with the target.
 - Statistical Methods
 - Feature Importance Methods
- **Wrapper:** Search for well-performing subsets of features.
 - RFE (Recursive Feature Elimination)
- **Intrinsic:** Algorithms that perform automatic feature selection during training.
 - Decision Trees



Filter:

- **Statistical-based feature selection methods** involve evaluating the **relationship** between each input variable and the target variable using statistics and selecting those input variables that have the strongest relationship with the target variable.
- These methods can be fast and effective, although the choice of **statistical measures depends on** the data type of both the **input and output variables**.



Wrapper:

- Wrapper feature selection methods **create many models with different subsets of input features** and **select those features that result in the best performing model according to a performance metric**

Intrinsic

- machine learning algorithms that perform feature selection automatically as part of learning the model. These techniques considered as **intrinsic** feature selection methods.
- **E.g. Decision Tree**



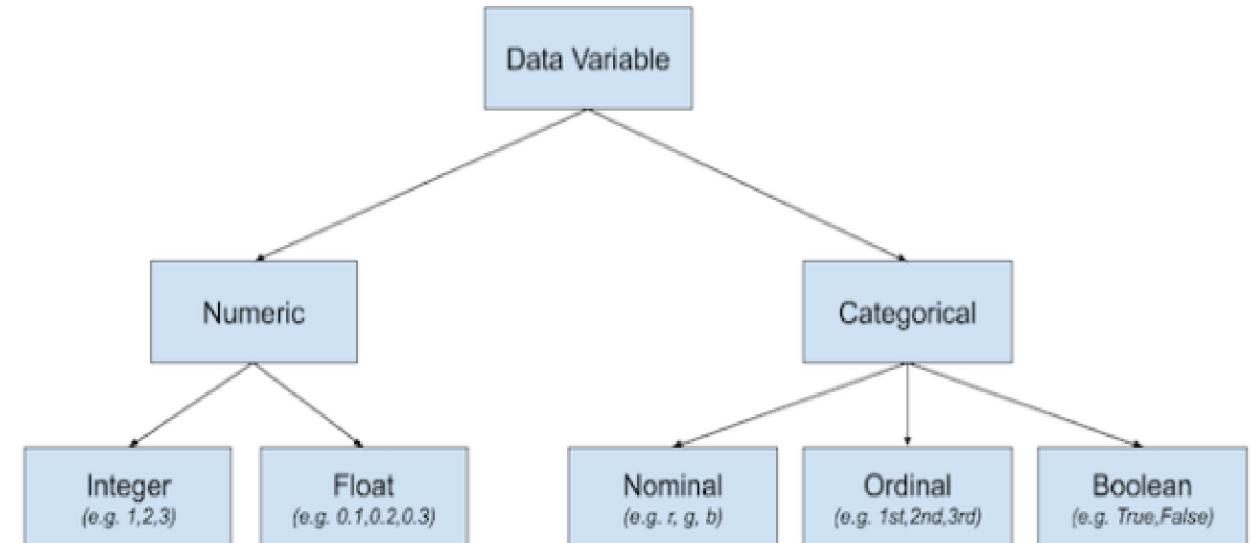
Statistics for Filter-Based Feature Selection Methods

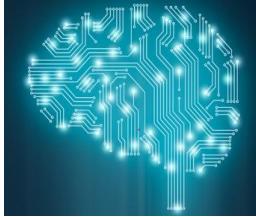
It is common to **use correlation** type statistical measures between input and output variables as the basis for filter feature selection.

Common input variable data types:

- **Numerical Variables**
 - Integer Variables.
 - Floating Point Variables.
- **Categorical Variables.**
 - Boolean Variables
 - Ordinal Variables.
 - Nominal Variables.

Overview of Data Variable Types





Feature Engineering: Feature selection

filter-based feature selection method.

- consider two broad **categories of variable types**: **numerical and categorical**
- Also the two main groups of **variables** to consider: **input and output**.
- **Input variables** are those that are provided as input to a model.
- In feature selection, it is this group of variables that we wish to reduce in size.
- **Output variables** are those for which a model is intended to predict, often called the response variable.

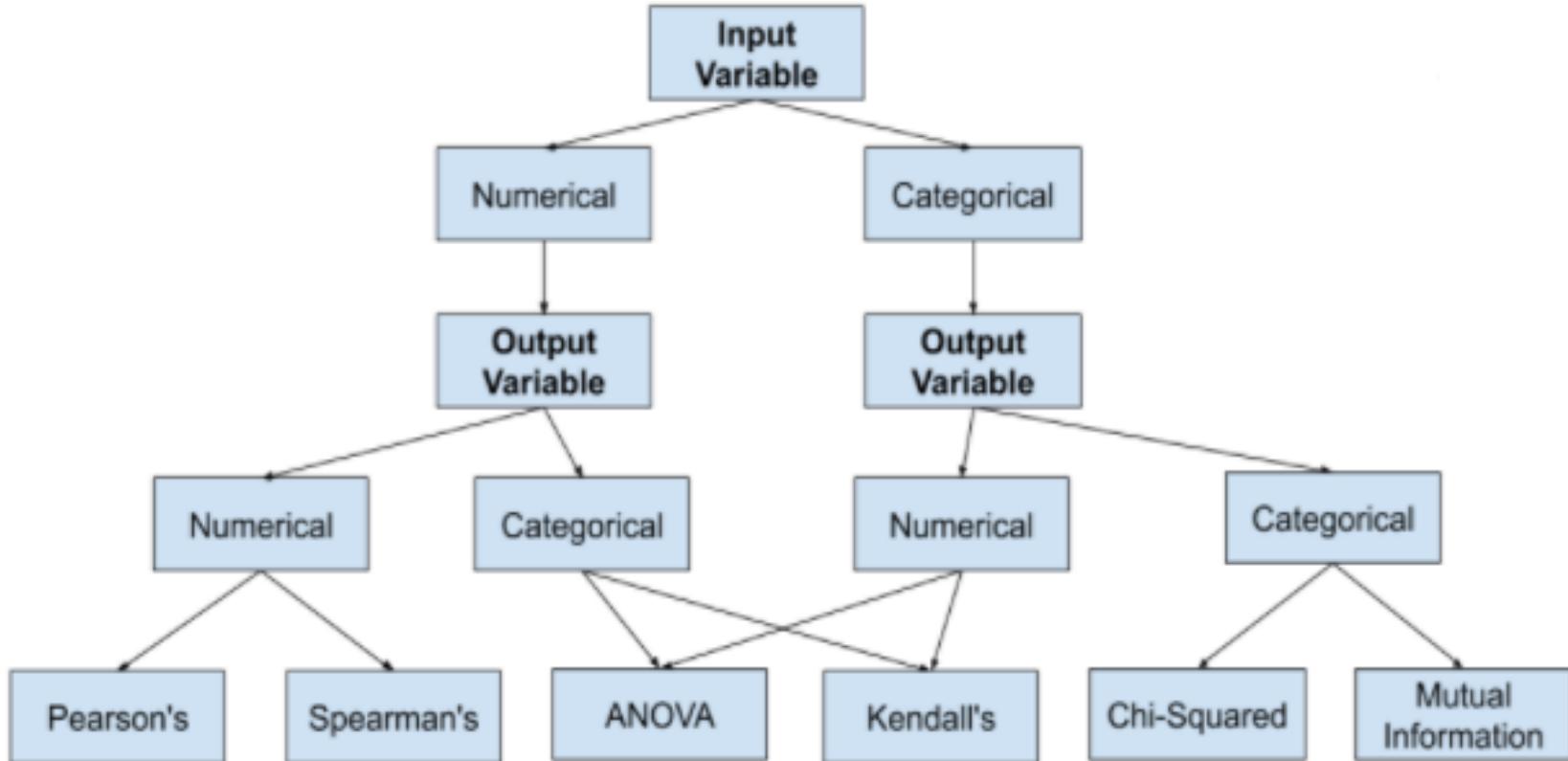
Response variable typically indicates the type of predictive modeling problem

- **Numerical Output:** **Regression** predictive modeling problem.
- **Categorical Output:** **Classification** predictive modeling problem



Feature Engineering: Feature selection

How to Choose a Feature Selection Method





1. Numerical Input, Numerical Output

- This is a **regression predictive modeling problem** with **numerical input variables**.
- The most common techniques are to use a correlation coefficient, such as Pearson's for a linear correlation, or rank-based methods for a nonlinear correlation.
- **Pearson's correlation coefficient (linear)**.
- **Spearman's rank coefficient (nonlinear)**

2. Numerical Input, Categorical Output

- This is a **classification predictive modeling problem** with **numerical input variables**.
- This might be the most common example of a classification problem,
- **ANOVA correlation coefficient (linear)**.
- **Kendall's rank coefficient (nonlinear)**.

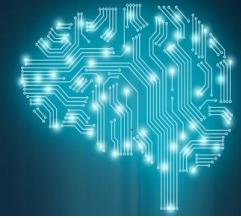


3. Categorical Input, Numerical Output

- This is a **regression predictive modeling problem** with **categorical input variables**.
- Nevertheless, you can use the same “*Numerical Input, Categorical Output*” but in reverse.

4. Categorical Input, Categorical Output

- This is a **classification predictive modeling problem** with **categorical input variables**.
- The most common correlation measure for categorical data is the [chi-squared test](#).
- Chi-Squared test (contingency tables).
- Mutual Information.



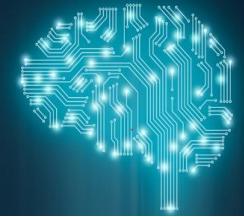
Recursive Feature Elimination

- Recursive Feature Elimination, or **RFE** for short, is a feature selection algorithm.
- A machine learning dataset for classification or regression is comprised of rows and columns, like an excel spreadsheet.
- **Rows** are often referred to as **samples** and **columns** are referred to as **features**
- Feature selection refers to techniques that **select a subset of the most relevant features (columns) for a dataset**.
- **Fewer features** can allow machine learning algorithms to **run more efficiently** (less space or time complexity) and be more **effective**.
- Some **machine learning algorithms** can be **misled by irrelevant input features**, resulting in worse predictive performance.

Feature Engineering: Feature selection



- RFE is a wrapper-type feature selection algorithm.
- This means that a **different machine learning algorithm** is given and used in the core of the method, is **wrapped by RFE**, and **used to help select features**.
- This is **in contrast** to filter-based feature selections that **score each feature** and **select those features with the largest (or smallest) score**.
- RFE works by **searching for a subset of features** by starting with all features in the training dataset and successfully removing features until the desired number remains.



Dimensionality Reduction: PCA

- We can see/visualize 2D,3D databy scatterplot
- 4D, 5D, 6Duse pair plot.....($nC2$ pairs)
- 10D,100D,1000D data ?
- Visualization of High dimension (n- dim) data?

n-D Reduce 2-D or 3-D

- **map high-dimensional data into low dimensions** and preserve all the structure.
- Using Dimensionality Reduction techniques like PCA, t-SNE to visualize high dimension data
- PCA tries to preserve linear structure, MDS tries to preserve global geometry, and t-SNE tries to preserve topology



MIT-WPU

॥ विद्यानन्तर्धावं ध्रुवा ॥



Principal Component Analysis (PCA)

Why PCA ?

- For dimensionality reduction i.e. d -dim $\rightarrow d'$ -dim E.g. mnist dataset of 784- dim to 2 dim

MNIST dataset downloaded from Kaggle :

#<https://www.kaggle.com/c/digit-recognizer/data>

Application-

- Visualization of high dim data using scatter plot, pair plot etc.
- As a ML ,models to solve problems on high dimensions



MIT-WPU

॥ विद्यानन्तर्धावं ध्रुवा ॥



Principal Component Analysis (PCA)

PCA Steps for dimensionality reduction:

1. Column standardization of data
2. Find Covariance matrix
3. Find eigen values and eigen vectors
4. Find principal components
5. Reducing dimensions of dataset



MIT-WPU

॥ विद्यानन्तर्धावं ध्रुवा ॥

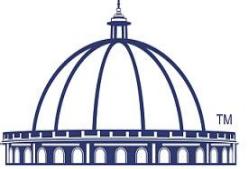


Principal Component Analysis (PCA)

1. Standardization of the data

- missing out on standardization will probably result in **a biased outcome**.
- Standardization is all about **scaling** your data in such a way that all the **variables and their values lie within a similar range**
- E.g let's say that we have 2 variables in our data set, one has values ranging between 10-100 and the other has values between 1000-5000.
- In such a scenario, it is obvious that the output calculated by using these predictor variables is going to be biased
- standardizing the data into a comparable range is very important.

$$Z = \frac{\text{Variable value} - \text{mean}}{\text{Standard deviation}}$$



Principal Component Analysis (PCA)

2 Computing the covariance matrix

- A covariance matrix expresses the **correlation between the different variables in the data set**.
- It is essential to identify **heavily dependent variables** because they **contain biased and redundant information which reduces the overall performance of the model**.
- a covariance matrix is a $p \times p$ matrix, where p represents the dimensions of the data set.
- 2-Dimensional data set with **variables a and b**, the covariance matrix is a 2×2 matrix as shown below

$$\begin{bmatrix} \text{Cov}(a, a) & \text{Cov}(a, b) \\ \text{Cov}(b, a) & \text{Cov}(b, b) \end{bmatrix}$$

- If the covariance value is negative, it denotes the respective variables are indirectly proportional to each other
- A positive covariance denotes that the respective variables are directly proportional to each other



Principal Component Analysis (PCA)

3. Calculating the Eigenvectors and Eigenvalues

Eigenvectors and eigenvalues are computed **from the covariance matrix** in order **to determine the principal components of the data set.**

What are Principal Components?

- Principal components are the new set of variables that are **obtained from the initial set of variables**.
- The principal components are computed in such a manner that newly obtained variables are **highly significant and independent of each other**.
- The principal components compress and **possess most of the useful information** that was scattered among the initial variables.
- E.g data set is of 5 dimensions, then 5 principal components are computed, such that, the first principal component stores the maximum possible information and the second one stores the remaining maximum info and so on



Principal Component Analysis (PCA)

4. Computing the Principal Components

- Eigenvectors and eigenvalues placed in the descending order
- where the **eigenvector with the highest eigenvalue is the most significant** and thus forms the **first principal component**.
- The principal components of **lesser significances can thus be removed** in order **to reduce the dimensions of the data**.
- The **final step in computing the Principal Components is to form a matrix known as the feature matrix** that **contains all the significant data variables that possess maximum information about the data**.



MIT-WPU

॥ विद्यान्तर्घुवं ध्रुवा ॥

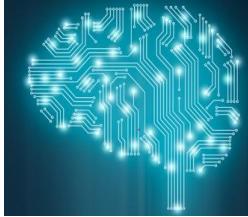


Principal Component Analysis (PCA)

5. Reducing the dimensions of the data set

- performing PCA is to **re-arrange the original data** with the **final principal components** which represent the maximum and the most significant information of the data set.
- In order to **replace the original data axis with the newly formed Principal Components**, you simply multiply the transpose of the original data set by the transpose of the obtained feature vector.
- For iris <https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60>

T-SNE



<https://distill.pub/2016/misread-tsne/>

<https://colah.github.io/posts/2014-10-Visualizing-MNIST/>

t-SNE is t-distributed stochastic neighborhood embedding

- Used dimensionality reduction
- Best technique for visualization
- PCA and t-SNE used in industry
- PCA preserve global structure whereas t-SNE preserve local structure



T-SNE.

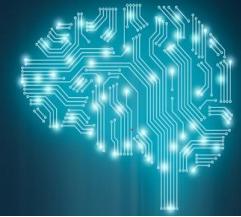
Neighborhood and Embedding

- Points are geometrically together.....Neighborhood
- Embedding....For every points in high-dim space finding its corresponding points in low dimension

Stochasticprobabilistic

Geometric intuition...preserving distances of points in neighborhood

T-SNE

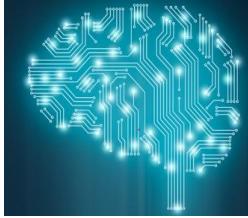


Crowding problem :

E.G ...2dim to 1 dim

Sometimes it is impossible to preserve distance in all the neighborhood points such problem is called crowding problem.

T-SNE

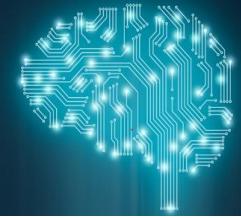


<https://distill.pub/2016/misread-tsne/>

- Run t-SNE on simple dataset
- Perplexity : points in neighbors
- Epsilon : learning rate
- Steps : iteration

t-SNE is iterative algorithm...1,2,3.... Runs t-SNE eventually till points are not moving (stable configuration...shape does not changing)

1. Always runs t-SNE till shape does not change
2. Always runs t-SNE with multiple perplexity value.
3. Perplexity $2 \leq p \leq N$



Exploratory Data Analysis

In statistics, exploratory data analysis (EDA) is an approach to **analyzing data sets to summarize their main characteristics, often with visual methods.**

- Understanding dataset by **using tools** from statistics or simple plotting tools
- Understand data ?
- E.G iris dataset....flowers are visually similar
- https://en.wikipedia.org/wiki/Iris_flower_data_set

A close-up photograph of two people in business attire shaking hands. The person on the left is wearing a dark suit and a white shirt cuff. The person on the right is wearing a light-colored suit and a white shirt cuff. The background is blurred, showing what appears to be a city street at night with some lights.

Thank you